Integrating Language Technologies and Social Theories

Anjalie Field

Thesis Committee
Yulia Tsvetkov (chair), Carnegie Mellon University
Alan Black, Carnegie Mellon University
Alexandra Chouldechova, Carnegie Mellon University
Dan Jurafsky, Stanford University

Submitted in partial conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Language Technologies Institute
Carnegie Mellon University

Abstract

In recent years, natural language processing (NLP) has seen rapid advancements over standardized tasks and carefully curated data sets. However, these models and benchmarks often fail to generalize to the diverse types of data and questions prevalent in society. In this thesis, I aim to develop technology capable of addressing diverse social-oriented questions in text by integrating social theories from related disciplines into NLP models. This work spans five primary social phenomena: stereotypes and prejudice in narrative text, global opinion manipulation strategies, toxicity on social media, public policy, and AI ethics. For each domain, I develop NLP models and frameworks that are grounded in relevant theories from other disciplines, including social psychology, political science, causal inference, and fairness. These methods are designed to involve minimal text annotations, instead relying on unsupervised or distantly supervised approaches, and several of them are language-agnostic or supported by cross-lingual models in order to facilitate analyses of languages beyond English, including Russian, Spanish, and Hindi. Overall, this thesis aims to shift NLP research beyond standard tasks and data sets to real-word data and challenges, where research questions and methodology are guided by relevant theories and incorporate social context.

Contents

Chapter 1

Introduction

The increasing availability of online text and platforms that disseminate information rapidly, like social media, online news, and Wikipedia, has amplified the proliferation of text content. Even in offline settings, text data is increasingly stored in machine-readable digitized format; for example, doctors' notes are now stored as electronic health records. The availability of online and digitized data is providing new opportunities for studying social phenomena that manifest in language (Lazer et al., 2009). Additionally, it has facilitated shifts further and further towards data-driven methodology in natural language processing (NLP). In the 1900s, rule-based methods gave way to statistics-based approaches. More recently, neural networks became the dominant paradigm, and even more recently pre-trained language models have dominated (Bommasani et al., 2021). These models require increasing amounts of text data to achieve high performance.

While publicly available digitized data has created new opportunities, at the same time, it has accelerated the spread of harmful content like hate speech and misinformation in ways that are difficult counter through human moderation. Language can be weaponized to manipulate public opinion (McCombs, 2002; Entman, 2007), and modern manipulation strategies have been termed "a critical threat to democracy" (Bradshaw et al., 2021). Language is also the primary means through which stereotypes and prejudices are communicated and perpetuated (Hamilton and Trolier, 1986; Bar-Tal et al., 2013), whose negative impacts have been well-documented (Krieger, 1990; Goldin, 1990; Steele and Aronson, 1995; Logel et al., 2009). Furthermore, a plethora of research has shown that NLP models are prone to absorbing and amplifying biases in training data (Bolukbasi et al., 2016; Zhao et al., 2017; Mora-Cantallops et al., 2019; Redi et al., 2020), and the large volume of data needed to train modern models makes it difficult to track data characteristics and quality (Bender et al., 2021). The rapid advancement in NLP model capabilities and increasing amount of unmonitored data has contributed to an ethical crisis: models trained on data without considering whom it describes, whom it was written by, and whom model outputs might affect are liable to amplify stereotypes, spread misinformation, and perpetuate discrimination.

However, the tendency of NLP models to absorb and amplify this type of content suggests that they are capable of detecting and characterizing it. This methodology could contribute to countering social issues like prejudice, stereotyping, and propaganda, through improving text quality and informing public policy as well as provide tools and guidance for mitigating harmful behavior in NLP systems. Significant challenges remain in developing methods to accomplish these tasks:

although NLP has seen rapid advancements in recent years, most NLP models are evaluated and developed over standardized tasks and carefully curated data sets that do not generalize to the diverse types of data and questions prevalent in society. We cannot answer questions like "is this newspaper outlet biased?" or "is this tweet likely to incite violence?" using a supervised classification model. Leveraging NLP technology to analyze social-oriented questions requires structuring broad research questions into concrete frameworks and developing NLP models capable of addressing them.

In this thesis, I aim to develop these frameworks and models by drawing on theories established through case studies and manual analyses in other fields, including psychology, political science, and journalism. Although the scale of text data and the prevalence of NLP models has greatly increased, the underlying social issues are not new, and we can draw from existing research in other fields. The methodology developed in this work facilitates uncovering and analyzing social phenomena at-scale using the computational power of NLP. It also results in both NLP technology that is useful for social science research (e.g., analyzing bias in society), as well as NLP models that are more robust to social context (e.g., analyzing bias in NLP models).

Developing NLP models for social-oriented task requires addressing several technical challenges that persist across applications domains. First, obtaining large annotated data can be difficult or impossible for many social phenomena and research questions. Concepts like "bias" are difficult to define or derive annotation schemes for (Blodgett et al., 2020), and it is challenging even for expert annotators to prevent their own biases from influencing the annotations. Furthermore, annotations are often specific to tasks and data sets, meaning they cannot be reused in other scenarios, especially because social concepts differ across contexts and cultures (Dong et al., 2019). Thus, I focus on developing distant supervision and domain adaptation methods that allow general annotations to be useful for task-specific questions.

Additionally, several aspects of this work highlight the need to adjust models for confounding variables and and imperfect measurements. Classic evaluation of NLP models focuses on numeric performance metrics, without requiring deep investigation into exactly what patterns models learn. In social-oriented domains, relying on models that learn correlative patterns or are trained on proxy variables can lead to incorrect conclusions or provide biased guidance. Methodology developed in this work to reduce these problems includes explicit demotion of confounding variables (Field and Tsvetkov, 2020; Field et al., 2022), benchmarking against external indicators (Field et al., 2018), and relying on technology to supplement or guide manual analyses, rather than full automation (Park et al., 2021; Tyagi et al., 2020; Field et al., 2022).

Overall this thesis aims to expand how we define and conceptualize NLP tasks in order to address broad social issues that manifest in text, but can be hard to define or identify. Rather than bringing complex issues into NLP research through curated data sets and well-defined benchmark tasks, this work aims to bring NLP research into society, through integration of existing social theory, development of methods that incorporate context, and consideration of ethical issues surround NLP technology.

Overview This thesis will be organized according to the application domains and social science concepts that provide frameworks and research questions:

- Chapter 2: Measuring Stereotypes and Prejudice in Narrative Text

  Application domain: Journalism and encyclopedias

Primary theory origins: Social psychology (Osgood et al., 1957; Russell, 1980, 2003)

Relevant papers: Field et al. (2019), Field and Tsvetkov (2019), Park et al. (2021), Field et al. (2022)

- Chapter 3: Uncovering Global Opinion Manipulation Strategies

  Application domain: Information integrity

  Theory origins: Political science (Entman, 2007; McCombs, 2002; Boydstun et al., 2013)

  Relevant papers: Field et al. (2018), Tyagi et al. (2020)

- Chapter 4: Identifying Toxicity on Social Media

  Application domain: Online incivility

  Theory origins: Causal inference (Rosenbaum and Rubin, 1983, 1985)

  Relevant papers: Field and Tsvetkov (2020), Xia et al. (2020)

- Chapter 5: Supporting Policy Decisions through Text Analysis

  Application domain: Public Policy

  Theory origins: Social psychology (Jasper, 2011; Goodwin et al., 2007; Allen and Leach, 2018), fairness and decision science (Brown et al., 2019; Coston et al., 2020)

  Relevant papers: Field et al. (Forthcoming(b)), Field et al. (Forthcoming(a))

- Chapter 6: Case Examination of the Risks and Harms of NLP

  Application domain and theory origins: NLP Research, AI Ethics

  Relevant papers: Field et al. (2021)

Following the main content chapters, §7 discusses the ethical implications considered in this work, and §8 concludes by highlighting the themes common across application domains as areas for future work.

Thesis Statement My goal is to develop technology capable of addressing diverse social-oriented questions that arise around text. My primary methodology involves devising applicable NLP models and frameworks that incorporate social context and are based on theories from related disciplines. I demonstrate the capabilities of this approach and how it results both in methodological contributions to addressing social issues and more socially-conscious NLP technology in several application domains. These domains include analyzing manifestations of stereotypes and prejudice, exposing opinion manipulation strategies, identifying online toxicity, and investigating text to inform public policy. This work is highly interdisciplinary, both in applications and methodology, as it draws theories and frameworks from social psychology, political science, and causal inference, demonstrating how these fields can inform the development of social-oriented NLP models. It also prioritizes the use of real-world data, and thus involves collecting and processing data from a variety of platforms and domains including Twitter, Facebook, Wikipedia, news articles, and expert-written notes about child welfare cases. While most data is in English, several parts involve analyses of data in other languages, including Russian, Hindi/Urdu, and Spanish. In developing NLP methodology for various domains, data types, and languages, this work broadly expands the conceptualization of NLP tasks and methods as well as the the usability of NLP technology in social domains.

Chapter 2

Measuring Stereotypes and Prejudice in Narrative Text

This chapter discusses work previously published in Field et al. (2019), Field and Tsvetkov (2019), Park et al. (2021), and Field et al. (2022).

Stereotypes and prejudices frequently manifest in text (Hamilton and Trolier, 1986; Bar-Tal et al., 2013), including text perpetuated as objective like newspaper articles and encyclopedias (Wagner et al., 2015). These social biases have the potential to influence readers and perpetuate well-documented negative impacts (Krieger, 1990; Goldin, 1990; Steele and Aronson, 1995; Logel et al., 2009). Furthermore, they are prone to be absorbed and amplified in NLP models (Bolukbasi et al., 2016; Zhao et al., 2017). The widespread adoption of language models pre-trained on large amounts of web data has drawn attention to biases in representations learned by these models (Kurita et al., 2019; Basta et al., 2019; Bender et al., 2021), though even models trained for specific tasks, often on carefully curated NLP datasets, also exhibit performance gaps and evidence of stereotypical associations (Webster et al., 2018; Rudinger et al., 2018; Zhao et al., 2019; Stanovsky et al., 2019).

Identifying signs of prejudice and stereotypes in narrative text has at least two direct applications: aiding authors in revising text and modifying NLP training data. In text like newspaper and encyclopedia articles, biased content is often unintentional. For example, Wikipedia maintains a "Neutral Point of View" policy,[1] and editors constantly revise and remove subjective content that violates this policy (Pryzant et al., 2020). In the past, research on other types social biases in Wikipedia, such as missing articles about women, has drawn the attention of the editor community and led to changes on the platform (Reagle and Rhue, 2011). Thus, automated tools for identifying signs of prejudice like content disparities, can aid authors in correcting them. Additionally, much work in NLP on "de-biasing" models has focused on correcting model representations or outputs after training (Bolukbasi et al., 2016; Zhao et al., 2017; Blodgett et al., 2020), but augmenting or balancing training data may be more effective at mitigating bias (Zhao et al., 2019). Tools for identifying content that is likely to result in biased models can proactively aid researchers in removing this content from model training data.

However, concepts like "social bias", "stereotypes", and "prejudice" can be difficult to define and even more difficult to identify. In this chapter, we infer that these concepts are all *people-centric*

[1] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

and fundamentally involve analyzing how people are described. Thus, we develop NLP models for measuring how people are portrayed in narrative text. In order to structure our approach, we draw from social psychology research, which has identified 3 primary dimensions as most important for capturing affective meaning: *Power* (strength vs. weakness), *Sentiment* (goodness vs. badness), and *Agency* (liveliness vs. torpidity) (Osgood et al., 1957; Russell, 1980, 2003).[2] These dimensions are considered both distinct and broad in capturing affective meaning, in that all 3 dimensions are needed, and that they capture primary affective meaning without needing additional dimensions; most other affective concepts, such as anger or joy, are thought to decompose into these three dimensions (Russell, 1980, 2003). Furthermore, these dimensions form the basis of *affective control theory*, a social psychological model which broadly addresses how people respond emotionally to events and how they attribute qualities to themselves and others (Heise, 1979, 2007; Robinson et al., 2006). Affective control theory has served as a model for stereotype detection in NLP (Joseph et al., 2017).

This chapter introduces *contextual affective analysis*, a framework for analyzing how people are described along dimensions of power, agency, and sentiment that enables nuanced, fine-grained, and directed analyses of affective social meanings in narratives. We first present a verb-centric approach, which we use to analyze online media coverage of the #MeToo movement, finding that while the media generally portrays women revealing stories of harassment positively, these women are often not portrayed as having high power or agency, which threatens to undermine the goals of the movement (§2.1). We then present an entity-centric approach and analyze general media coverage of specific public figures (§2.2). Finally, §2.3 discusses some of the additional methodology needed to analyze broader manifestations of stereotypes, specifically the need to control for confounding variables, and presents analyses of Wikipedia data. Overall, this chapter focuses on developing generalizable methodology to uncover manifestations stereotypes and prejudice in narrative text and presents use cases of these methods in various settings.

2.1 Verb-centric Contextual Affective Analysis

2.1.1 Background

We motivate the development of contextual affective analysis as a people-centric approach to analyzing narratives. Entity-centric models, which focus on people or characters rather than plot or events, have become increasingly common in NLP (Bamman, 2015). However, most prior approaches rely on unsupervised models (Chambers and Jurafsky, 2009; Bamman et al., 2013; Iyyer et al., 2016; Card et al., 2016), which can capture high-level patterns but are difficult to interpret and do not target specific dimensions.

In contrast, we propose an interpretable approach that focuses on power, sentiment, and agency. The crux of our method is in developing contextualized, entity-centric connotation frames, where polarity scores are generated for words in context, and supervised learning propagates annotations to unlabeled analysis corpora. While automated sentiment analysis has spanned many areas (Pang and Lee, 2008; Liu, 2012) analysis of power has been almost entirely limited to a dialog setting: how

[2] Exact terminology for these terms has varied across studies, e.g. Power has also been called Dominance or Potency. We use the terminology most common in NLP literature (Sap et al., 2017; Rashkin et al., 2016; Field and Tsvetkov, 2019; Field et al., 2019; Park et al., 2021)

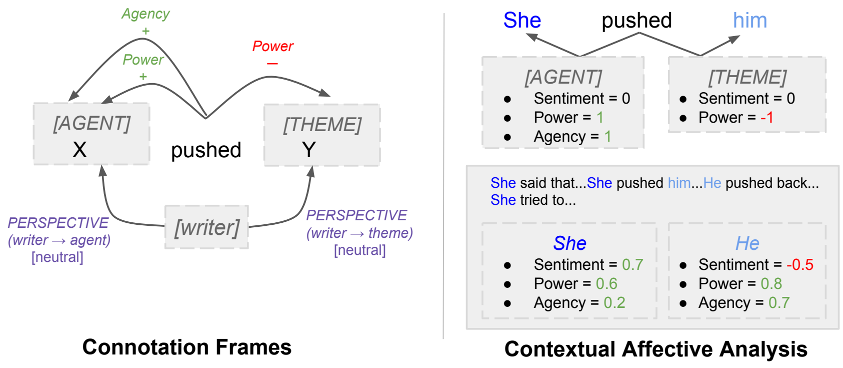


Figure 2.1: Left, we show off-the-shelf connotation frame annotations (Rashkin et al., 2016; Sap et al., 2017) for the verb "push". Right, we show the proposed adaptation. We adapt connotation frames from a verb-centric formalism to an entity-centric formalism, transferring scores from verbs to entities using a context-aware approach (top right). We then aggregate contextualized scores over all mentions of entities in a corpus (bottom right). This new approach—*contextual affective analysis*—enables us to obtain sentiment, power, and agency scores for entities in unannotated corpora and conduct extensive analyses of people portrayals in narratives, which we exemplify on #MeToo data.

does person A talk to a higher-powered person B? (Gilbert, 2012; Danescu-Niculescu-Mizil et al., 2012; Prabhakaran and Rambow, 2017). Here, we focus on a *narrative* setting: does the writer portray person A or person B as more powerful?

In order to develop an interpretable analysis that focuses on sentiment, power, and agency in narrative, we draw from existing literature on *connotation frames*: sets of verbs annotated according to what they imply about semantically dependent entities. Connotation frames, first introduced by Rashkin et al. (2016), provide a framework for analyzing nuanced dimensions in text by combining polarity annotations with frame semantics (Fillmore, 1982). We visualize connotation frames in Figure 2.1 on the left. More specifically, verbs are annotated across various dimensions and perspectives, so that a verb might elicit a positive sentiment for its subject (i.e. sympathy) but imply a negative effect for its object. We target power, agency, and sentiment of entities through pre-collected sets of verbs that have been annotated for these traits:

- Perspective($writer \rightarrow agent$) – Does the writer portray the agent (or subject) of the verb as positive or negative?
- Perspective($writer \rightarrow theme$) – Does the writer portray the theme (or object) of the verb as positive or negative?
- Power – does the verb imply that the theme or the agent has power?
- Agency – does the verb imply that the subject has positive agency or negative agency?

For clarity, we refer to Perspective($writer \rightarrow agent$) as Sentiment($agent$) and Perspective($writer \rightarrow theme$) as Sentiment($theme$) throughout this chapter.

These dimensions often differ for the same verb. For example, in the sentence: "She amuses him," the verb "amuses" connotes that *she* has high agency, but *he* has higher power than *she*. Rashkin et al. (2016) present a set of verbs annotated for sentiment, while Sap et al. (2017) present a set of verbs annotated for power and agency. However, these lexicons are not extensive enough to facilitate corpus analysis without further refinements. First, they contain only a limited set of verbs, so a given corpus may contain many verbs that are not annotated. Furthermore, verbs are annotated in synthetic context (e.g., "X amuses Y"), rather than using real world examples. Finally, each verb is annotated with a single score for each dimension, but in practice, verbs can have different connotations in different contexts.

Consider two instances of the verb "deserve":

1. The hero deserves appellation

2. The boy deserves punishment

In the first instance, annotators rate the writer's perspective towards the agent ("hero") as positive, while in the second instance, annotators rate the writer's perspective towards the agent ("boy") as negative (Rashkin et al., 2016). We can find numerous similar real-world examples in news articles i.e. *She pushed him away* vs *Will one part of the movement's legacy be to push society to find the right words to describe it all?*.

The uncontextualized annotations presented by Rashkin et al. (2016) and Sap et al. (2017) serve as starting points for more in-depth analysis. We build uncontextualized features for verbs to match the uncontextualized annotations, and then use supervised learning to extend the uncontextualized annotations to verbs in context, thus learning contextualized annotations.

Our work extends the concept of domain-specific lexicons: that words have different connotations in different situations. For instance, over the last century, the word "lean" has lost its negative association with "weakness" and instead become positively associated with concepts like "fitness." These changes in meaning have motivated research on inducing domain-specific lexicons (Hamilton et al., 2016). Using contextual embeddings (Peters et al., 2018), we extend this concept by introducing *context-specific* lexicons: we induce annotations for words in context, rather than just words in domain. To the best of our knowledge, this is the first work to propose contextualized affective lexicons. Our overall methodology uses these contextualized lexicons to obtain power, sentiment, and agency scores for entities in contextual affective analysis.

2.1.2 Methodology

In the proposed contextual affective analysis, our primary goal is to analyze how people are portrayed in narratives. To obtain sentiment, power, and agency scores for people in the context of a narrative (a sentence, paragraph, article, or an outlet), we adapt the connotation frames reviewed above as follows. First, since connotation frames are verb-centric—rather than people-centric—we define a mapping from a verb in a sentence to people that are syntactic arguments of the verb. We visualize this mapping in Figure 2.1. Second, since only a small subset of verbs (<30%) in our corpus is included in the annotations crowdsourced by Rashkin et al. (2016) and Sap et al. (2017), we devise a lexicon induction method to annotate unlabeled verbs using the existing seed of annotations. To contextualize the analysis, the lexicon induction procedure uses contextual ELMo embeddings (Peters et al., 2018). Contextual embeddings are a form of distributed word representations that incorporate surrounding context words. Thus, using the above example, ELMo embeddings provide different representations for "push" in "She pushed him away" and for "push" in "Will one part of the movement's legacy be to push..." In what follows, we detail the components of our methodology.

We use connotation frames, which are annotations on verbs, to obtain affective scores on people (the agent or theme of these verbs). In our running example "She pushes him away," "She" is the agent and "him" is the theme. Given a verb V, the verb's agent A, the verb's theme T, and a set of connotation frame annotations over V (e.g., $V_{Sentiment(agent)} = +1$), we obtain sentiment, power, and agency scores as follows:

$$\begin{aligned} Sentiment(A) &= V_{Sentiment(agent)} \\ Power(A) &= V_{Power} \\ Agency(A) &= V_{Agency} \\ Sentiment(T) &= V_{Sentiment(theme)} \\ Power(T) &= -V_{Power} \end{aligned}$$

We obtain a sentiment score within a corpus for an entity E by averaging over all $V_{Sentiment(agent)}$ scores where E is the agent of V and all $V_{Sentiment(theme)}$ scores where E is the theme of V. We compute agency and power scores in the same way. Thus, the final entity scores are not merely direct mappings from verb annotations, but an aggregation of all such mappings over the corpus of entity mentions.

We obtain these verb scores by taking a supervised approach to labeling verbs in context with sentiment ($V_{Sentiment(agent/theme)} \in \{-1, 0, +1\}$), power, ($V_{Power} \in \{-1, 0, +1\}$), and agency ($V_{Agency} \in \{-1, 0, +1\}$).

For a given verb in our training set V, we assume that V occurs n times in our corpus, and enumerate these occurrences as $V_1 ... V_n$. For each V_i, we compute the contextualized ELMo embedding $\mathbf{e}_i$. We then "decontextualize" these embeddings by averaging over $\mathbf{e}_1 \dots \mathbf{e}_n$ to obtain a single feature representation $\mathbf{e}$. We consider these decontextualized embeddings to be representative of the off-the-shelf uncontextualized lexicons, and we use them as training features in a supervised classifier.

Then, for a given verb in our corpus T, for each instance where T occurs in our corpus, we use its contextualized ELMo embedding $\mathbf{e}_i$ as a feature to predict an annotation score for T_i. In particular, we use logistic regression with re-weighting of samples to maximize for the best average F1 score over a dev set.[3]

For evaluation, in order to compare our results against existing annotations, we use two methods to obtain uncontextualized annotations from the learned T_i scores for verbs in our test sets. In the first, which we refer to as *type-level*, we average all of the token-level embeddings ($\mathbf{e}_i$) in the test data in the same way as in the training data, and we learn a single annotation for each verb T, rather than learning contextualized annotations. This approach is most similar to prior work. In the second, which we refer to as *token-level*, we predict a separate score for each token-level embedding as described above, and we then take a majority vote over scores to obtain an overall score for each verb.

[3] We experimented with other supervised and semi-supervised methods common in lexicon induction including graph-based semi-supervised label propagation and random walk-based propagation (Goldberg and Zhu, 2006; Hamilton et al., 2016), but found that logistic regression outperformed these methods on all frames.

	Lexicon Size	Training Set Size
Sentiment	948	300
Power	1,714	571
Agency	2,104	701

Table 2.1: Annotated Lexicon Statistics

	Aspect	Frame	Type	Token
Sentiment (*theme*)	56.18	56.18	55.63	51.90
Sentiment (*agent*)	60.72	63.07	58.26	60.23
		Majority Class	Type	Token
Power	-	27.37	55.97	54.10
Agency	-	29.45	48.79	50.14

Table 2.2: F1 scores of lexicon expansion for our methods (Type and Token) compared with prior work (Aspect and Frame). Non-trivial F1 scores demonstrate that our feature representations capture meaningful information about sentiment, power, and agency.

2.1.3 Evaluation

We first evaluate our methods on their ability to predict the labels of off-the-shelf connotation frame lexicons. While this task is not our primary objective, it serves as a sanity-check on our feature representations and allows us to compare our method with prior work. Then, we evaluate the methods in a contextualized setting, assessing their ability to model contextualized verb annotations and contextualized entity annotations by comparing with human annotators.

We provide data sizes of the annotated lexicons in Table 2.1 and refer to Rashkin et al. (2016) and Sap et al. (2017) for additional details. To obtain sentences containing lexicon verbs that we can use to extract ELMo embeddings, we use data from a corpus 27,602 online articles about the #MeToo movement, which we describe in more detail in §2.1.4. We divide the annotations into train, dev, and test; for sentiment, we use the same data splits as Rashkin et al. (2016); for power and agency, we randomly divide the lexicons into subsets of equal size.

Uncontextualized Evaluation In order to compare the contextualized scores generated by our method with the off-the-shelf annotations, we aggregate the contextualized annotations into uncontextualized annotations for the verbs in the test set, as described in §2.1.2. Table 2.2 reports results. For comparison, we show the Aspect-Level and Frame-Level models presented by Rashkin et al. (2016) over the sentiment annotations. Our type-level logistic regression is essentially identical to the aspect-level model, the primary difference being our use of ELMo embeddings. Our results are slightly lower, but generally comparable to the results reported by Rashkin et al. (2016); crucially, they are obtained with a model that ultimately allows us to incorporate context. The type-level and token-level aggregation methods perform about the same.

In the absence of prior work on this task for the power and agency lexicons, we show a majority class baseline. Our methods show a strong improvement over F1 scores for the majority class baseline. As for sentiment, the type-level and token-level methods perform similarly. Table 2.2 generally shows that ELMo embeddings capture meaningful information about power, agency, and sentiment. However, our primary task is not to re-create the word-level annotations in the connotation frame lexicons, but rather to contextualize these lexicons by obtaining instance-level scores over verbs in

	Verb-level	Sent.-level	Sent.-level training
Sentiment (t)	41.05	44.35	50.16
Sentiment (a)	51.37	52.80	54.11

Table 2.3: F1 scores for using our method to score contextualized annotations. Predicting sentence-level scores outperforms predicting verb-level scores. Best performance is achieved by also using sentence-level training data, but this can be difficult in practice.

context, which we evaluate next.

Contextualized Evaluation We draw from the original annotations used to create the connotation frame lexicons in order to assess the impact of contextualization. The publicized connotation frames consist of a single score for each verb. However, for the sentiment dimensions, these scores were obtained by collecting annotations over verbs in a variety of simple synthetically-generated contexts and averaging annotations across contexts, i.e. collecting 5 annotations each for "the hero deserves appellation," "the student deserves an opportunity," and "the boy deserves punishment" and averaging across the 15 annotations (Rashkin et al., 2016). Then, for the sentiment lexicons, we can evaluate our method's ability to provide annotations in context by reverting to the original pre-averaged annotations. (We cannot perform the same evaluation for the power and agency lexicons, because they were created by collecting annotations over verbs without any context, i.e. "X deserves Y" (Sap et al., 2017).)

When we ignore context, meaning we treat all 15 annotations over each verb as annotations over the same sample, the inter-annotator agreement (Krippendorff's alpha) for Sentiment(*theme*) is 0.20 and for Sentiment(*agent*) is 0.28. However, when we treat each sentence as a separate sample (i.e. measuring agreement in annotations over "the hero deserves appellation" separately from annotations over "the boy deserves punishment"), the agreement rises to 0.34 and 0.40 respectively, a $>40\%$ increase for each trait. The improvement in agreement suggests that when annotators disagree about the connotation implied by a verb, it is often because the verb has different connotations in different contexts.

We can then evaluate our method for contextualization by using these sentence-level annotations. More specifically, we use the same train, dev, and test splits as before. However, for verbs in the test set, instead of averaging all 15 annotations for each verb into a single score, we only average over annotations on the same sentence. Thus, our gold test data contains separate scores for "the hero deserves appellation", "the student deserves an opportunity", and "the boy deserves punishment", resulting in approximately 3 times as many test points as the test data in Table 2.2.

Table 2.3 shows the results of evaluating our method on this contextualized test set. In the first column, Verb-level, our model disregards contextualization and predicts a single score for each verb using type-level aggregation, which is equivalent to the method used in Table 2.2. The primary difference is that in Table 2.2, we evaluate over uncontextualized annotations (i.e. a single score for "deserve"), while in Table 2.3, we evaluate over contextualized annotations. In the second column, Sent.-level, we predict a separate score for each verb in context, rather than using token or type level aggregation over the test data. This column represents our primary method and is the method we use for analysis in §2.1.4. For both traits, this method outperforms the aggregation approach shown in the first column. In the third column, Sent.-level training, we similarly predict a separate score for each context, but we further treat each context as a separate training sample. Thus we both train and evaluate on contextualized annotations. In the Sent.-level and Verb-level columns, we train on uncontextualized annotations, as described in §2.1.2.

While training on contextualized annotations achieves the best performance, it is difficult to generalize to other data sets. The sentences used for gathering these annotations were created using Google Syntactic N-grams and designed to be short generic sentences (Rashkin et al., 2016). Thus, they are much simpler than real sentences, and we would not expect them to serve as realistic training data in other domains. In order to use these connotation frame lexicons in a new domain, it would be necessary to annotate a new set of sentences, which defeats the usefulness of off-the-shelf lexicons. Instead, we focus on the second column, Sent.-level, as our primary method, since it is an improvement over existing ways of using off-the-shelf lexicons without requiring new annotations for every task. Overall, Table 2.3 demonstrates the usefulness of contextualization, as both contextualized approaches outperform the uncontextualized approach.

Evaluation of Entity Scores In the previous paragraphs, we evaluated our methods for generating verb annotations. Here, we evaluate our methods for transferring verb annotations to entity scores, specifically focusing on power. In order to assess entity scoring, we devised an annotation task in which we asked annotators to read articles and rank entities mentioned. We then compare our entity scores against these annotations.

More specifically, we sampled 30 articles from our corpus that all mentioned the same person (Aziz Ansari). We then provided 2 annotators with a list of 23 entities extracted from these articles and asked the annotators to read each article in order. After every 5 articles, annotators ranked the listed entities by assigning a 1 to the lowest-powered entity, a 7 to the highest-powered entity, and scaling all other entities in between. In this way, since annotators rerank entities every 5 articles, we maximize the number of annotations we obtain, while minimizing the number of articles that annotators need to read. Furthermore, we ensure that we obtain different power scores for different entities by forcing annotators to use the full range of the ranking scale.

However, we note that this is a very subjective annotation task. Annotators specifically described it as difficult, both in deciding which entities were most powerful and in ranking entities based on the provided articles rather than on outside knowledge. We observed some of this subjectivity in the collected annotations: one annotator consistently ranked abstract entities like “The New York Times” as high-powered, while the other annotator consistently ranked these entities as low-powered. Thus, while we present results as an approximation of how well our methods work, we caution that further evaluation is needed.

From the annotations, we have a ranking for each entity for each 5-article step. Hypothesizing that some entities are more subjective to rank than others, we eliminate all samples where the difference in ranking between the two annotators is greater than 2. We are then left with 81 annotations. The correlation between annotators on this set is statistically significant (Spearman’s R=0.55, p-value=1.03e-07). In the following analysis, we average the rank assigned by annotators to obtain a score for each entity.

We ultimately evaluate our methods through pairwise comparisons at each 5-article step. For every pair (A, B) of entities at each time step, we evaluate if entity A is scored as more powerful or less powerful than entity B. We discard samples where the entities were ranked as equal. We

Off-the-shelf	Frequency	Ours
57.1	59.1	71.4

Table 2.4: Accuracy for scoring how powerful entities are, as compared with manual annotations. We calculate accuracy by assessing if the metric correctly answers "is entity A more powerful than entity B?". Our method outperforms both baseline metrics.

compare our method against 2 baseline metrics: (1) the frequency of the entity and (2) power scores assigned by the off-the-shelf connotation frames, rather than our contextualized frames.

Off-the-shelf connotation frames are limited to a subset of verbs in our corpus. Furthermore, our analysis pipeline is dependent on the named entity recognition and co-reference resolution tools used during pre-processing to extract entity mentions, which we find miss many mentions (for instance, when we manually extract entities in the first 10 articles, we identify 28 entities that occur at least 3 times, while the automated pipeline identifies only 4). For fair comparison between our method and the off-the-shelf lexicons, we discard entities that do not occur with at least 3 off-the-shelf power-annotated verbs, as identified by our preprocessing pipeline. After this filtering, we are left with 49 pairwise comparisons.

Table 2.4 reports results. Our method outperforms both baselines, correctly identifying the higher-ranked entity 71.4% of the time. We note that each annotator individually achieves at most 83.7% accuracy on this task, which suggests an upper limit on the achievable accuracy.

Furthermore, if we perform the same test, eliminating only entities that occur fewer than 3 times in the text, rather than mandating that entities occur with at least 3 off-the-shelf annotated verbs, our method achieves an accuracy of 63.01% over 73 pairs, while the frequency baseline achieves an accuracy of 53.42%.

In conducting this analysis, we observed that one of the limitations of the power annotations provided by Sap et al. (2017) is that the authors only annotate transitive verbs for power. They hypothesize that a power differential only occurs when an entity (e.g. the agent) has power over another entity (e.g. the theme). However, we do not limit our scoring to transitive verbs, hypothesizing that intransitive verbs can also be indicative of power, even if there is no direct theme. The improved performance of our scoring metric over the off-the-shelf baseline supports this hypothesis.

2.1.4 Analysis of Entity Portrayals in the #MeToo Movement

In this section, we demonstrate how contextual affective analysis facilities examination of narrative corpora by examining people portrayals media coverage of the #MeToo movement. Tarana Burke founded the #MeToo movement in 2006, aiming to promote hope and solidarity among women who have experienced sexual assault.[4] In October 2017, following waves of sexual harassment accusations against producer Harvey Weinstein, actress Alyssa Milano posted a tweet with the hashtag #MeToo and encouraged others to do the same. Her message initiated a widespread movement, calling attention to the prevalence of sexual harassment and encouraging women to share their stories.

Tarana Burke has described her primary goal in founding the movement as "empowerment through empathy."[5] However, mainstream media outlets vary in their coverage of these recent events,

[4] https://www.washingtonpost.com/news/the-intersect/wp/2017/10/19/the-woman-behind-me-too-knew-the-power-of-the-phrase-when-she-created-it-10-years-ago/

[5] https://metoomvmt.org/

to the extent that some outlets accuse others of misappropriating the movement. For instance, in January 2018, Babe.net published an article written by Katie Way, describing the interaction between anonymous 'Grace' and famous comedian Aziz Ansari.[6] The article sparked not only instant support for Grace, but also instant backlash criticizing Grace's lack of agency: "The single most distressing thing to me about this story is that the only person with any agency in the story seems to be Aziz Ansari".[7] One widely circulated article, written by Caitlin Flanagan and published in The Atlantic, strongly criticized Way's article and questioned whether modern conventions prepare women to fight back against potential abusers.[8] Thus, while our focal dimensions of power, agency, and sentiment have origins in social psychology, they are particularly relevant in the context of the #MeToo movement and tie closely to the concept of "empowerment through empathy."

The manner in which accounts of sexual harassment portray the people involved affects both the audience's reaction to the story and the way people involved in these incidents interpret or cope with their experiences (Spry, 1995). Unlike prior work focused on social media (Ribeiro et al., 2018; Rho et al., 2018), our work examines the prominent role that more traditional outlets and journalists continue to have in the modern-era online media landscape. To examine this data type, we gathered a corpus of articles related to the #MeToo movement by first collecting a list of URLs of articles that contain the word "metoo" using an API that searches for articles from over 30,000 news sources.[9] Over two separate queries, we gathered URLs from November 2, 2017 to January 31, 2018 and from February 28, 2018 to May 29, 2018. Next, we used Newspaper3k to obtain the full text of each article.[10] After data cleaning and de-duplication, our final data set consists of 27,602 articles across 1,576 outlets, published between November 2, 2017 and May 29, 2018. We note that our data set consists of an imperfect sample of articles covering the #MeToo movement and a different sampling of articles may yield different results. We used the Stanford NLP pipeline for dependency parsing, named entity recognition, and co-reference resolution in order to extract mentions of people and corresponding verbs from the text. We found a total of 3,132,389 entity-verb pairs across the corpus, which form the basis of our analysis. Field et al. (2019) provide additional details on data collection and preprocessing.

We use the methodology introduced in §2.1.2 to obtain power, agency, and sentiment scores for mentions of people in the corpus, and we propose a top-down framework for structuring a people-centric analysis with three primary levels:

1. Corpus-level: we examine broad trends in coverage of all common entities across the entire corpus

2. Role-level: we examine how people in similar roles across separate incidents are portrayed

3. Incident-level: we restrict our analysis to people involved in a specific incident

Corpus-Level By examining portrayals at a corpus level, we can assess the overall media coverage of the #MeToo movement. Whom does the media portray as sympathetic? Did media coverage of events empower individuals?

[6] https://babe.net/2018/01/13/aziz-ansari-28355
[7] https://www.nytimes.com/2018/01/15/opinion/aziz-ansari-babe-sexual-harassment.html
[8] https://www.theatlantic.com/entertainment/archive/2018/01/the-humiliation-of-aziz-ansari/550541/
[9] https://newsapi.org/
[10] http://newspaper.readthedocs.io/en/latest/

Positive	Kara Swisher, Tarana Burke, Meghan Markle, Frances McDormand, Oprah Winfrey
Negative	Bill Cosby, Harvey Weinstein, Eric Schneiderman, Kevin Spacey, Ryan Seacrest

Table 2.5: The most positively portrayed entities consist primarily of 3$^{\text{rd}}$ party commentators on events. The most negatively portrayed entities consist of men accused of sexual harassment.

Highest Power	**Lowest Power**	**Highest Agency**	**Lowest Agency**
The #MeToo movement	Kevin Spacey	Judge Steven O'Neill	Kara Swisher
Judge Steven O'Neill	Andrea Constand	Eric Schneiderman	the United States
The New York Times	Uma Thurman	Russell Simmons	Hollywood
Congress	Dylan Farrow	The New York Times	Meryl Streep
Facebook	Leeann Tweeden	Frances McDormand	
Twitter		CNN	
Eric Schneiderman		Donald Trump	
Donald Trump		Hillary Clinton	

Table 2.6: The entities portrayed as most powerful consist of men and abstract institutions. The lowest powered entities consist of primarily women. Entities with the most the agency and the least agency consist of a mix of men, women, and abstract institutions.

We examine these questions by computing sentiment, power, and agency scores for the 100 most frequent proper nouns across the corpus, shown in part in Tables 2.5–2.6. For brevity, we omit redundant entities (i.e. Donald Trump and President Donald Trump). Table 2.5 shows the five most positively and negatively portrayed entities. Unsurprisingly, the most negatively portrayed entities all consist of men accused of sexual harassment, led by Bill Cosby, the first man actually convicted in court following the wave of accusations in the movement. However, the most positively portrayed entities consist not of women voicing accusations or of men facing them, but rather of 3$^{\text{rd}}$ party commentators, who were outspoken in their support of the movement. None of these entities were directly involved in cases that arose out of the #MeToo movement, but all of them made widely-circulated comments in support of the accusers.

When we examine power (Table 2.6), the most powerful entities include abstract concepts, like "the #MeToo movement" and "Twitter". Women are conspicuously absent from the list of high-powered entities. Instead we find men, including ones directly accused of sexual misconduct (Eric Schniederman). In contrast, women dominate the list of lowest powered entities. While Table 2.6 only shows proper nouns, we also observed that common noun references to women were among the least powerful entities identified (e.g. "a women", "these women"). The agency portrayals are more balanced. While Eric Schneiderman and Donald Trump appear among the entities with highest agency, we also find female supporters of the movement: Frances McDormand and Hillary Clinton.

Role-Level We next conduct a pairwise analysis, where we directly compare two entities who occupied similar roles in different incidents. Through this analysis, we can identify different ways in which narratives of sexual harassment can be framed. Furthermore, by decomposing the pair-wise analysis across different outlets, we can identify bias. How do different outlets cover comparable entities differently?

We directly compare sentiment and power for several entity pairs (Figure 2.2). There is a striking difference between the portrayals of Rose McGowan and Leeann Tweeden, two women who both accused high-profile men of sexual assault (Harvey Weinstein and Al Franken respectively). Both

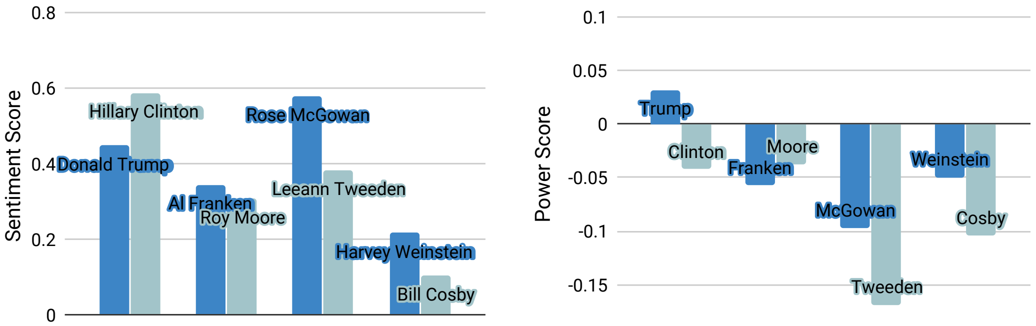


Figure 2.2: Entities in similar roles are portrayed with different levels of sentiment. Power scores for comparable entities do not coincide with sentiment scores.

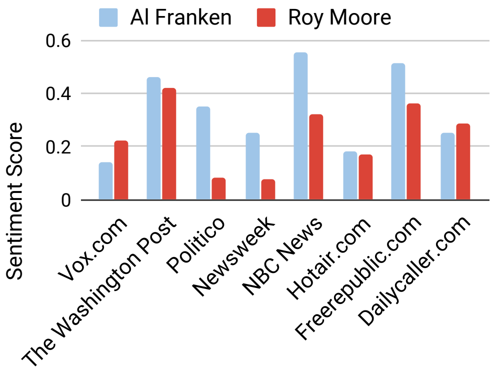


Figure 2.3: Sentiment scores across left-leaning and right leaning outlets for Democrat Al Franken and Republican Roy Moore do not fall along party lines.
Left-leaning (Democratic) outlets: Vox.com, The Washington Post, Newsweek, NBC News.
Right-leaning (Republican) outlets: Hotair.com, Freerepublic.com, Dailycaller.com.
Centrist: Politico
Outlet ideologies are taken from https://www.allsides.com/.

women are portrayed with lower power than the men they accused, but Rose McGowan is portrayed with both more positive sentiment and higher power than Leeann Tweeden.

Sample articles about Leeann Tweeden focus on accounts of what happened to her: "The first women to speak out was Leeann Tweeden who said that Franken forcibly kissed her."[11] In contrast, news articles about Rose McGowan focus on statements she made after the fact: "As Rose McGowan, one of the heroes to emerge from the Harvey Weinstein fallout, has mentioned...".[12] We can generalize that in this corpus, Rose McGowan matches more of "survivor" frame, connoting someone who is proactively fighting, while Leeann Tweeden matches more of a "victim" frame, which connotes helplessness and pity.

Figure 2.2 reveals further differences in portrayals. Democrats Hillary Clinton and Al Franken are portrayed more positively than corresponding Republicans Donald Trump and Roy Moore. However, Donald Trump appears as more powerful than every other entity, which coincides with his role as the current U.S. President. In general, politicians accused of sexual harassment (Roy Moore and Al

[11] https://dailym.ai/2WBiA38

[12] https://www.fastcompany.com/40513979/political-statements-black-dresses-and-surprise-wins-what-to-expect-at-the-golden-globes-this-sunday

Franken) are portrayed more positively than entertainment industry figures (Harvey Weinsten and Bill Cosby). While few defended entertainment industry figures accused of harassment, politicians received more mixed support and criticism from their own parties. For instance, Donald Trump publicly endorsed candidate Roy Moore, despite the allegations against him.

However, comparing coverage of individuals across the corpus as a whole reveals limited information because it is difficult to separate fact from bias. Does Al Franken receive a more positive portrayal than Roy Moore because his actions throughout the movement were more sympathetic? Or does he receive a more positive portrayal because of a liberal media bias? We can better assess the impact of media bias by comparing how the same entity is portrayed across different outlets. Figure 2.3 shows sentiment scores for Al Franken and Roy Moore across all outlets that mention both entities at least 10 times.

The coverage of Republican Roy Moore falls broadly along party lines, with the lowest-scoring portrayals occurring in left-leaning outlets Politico and Newsweek. Similarly, the left-leaning outlets (with the exception of Vox.com) portray Democrat Al Franken more positively than Roy Moore. However, the right-leaning outlets portray both men similarly, notably Freerepublic.com portrays Al Franken more positively than Roy Moore. In reading articles from these outlets, many are sympathetic towards Al Franken, presenting him as a scapegoat, forced out of office by other Democrats without a fair ethics hearing.[13] Our analysis reveals surprising differences in how conservative and liberal outlets react differently to events of the #MeToo movement. Entity portrayals do not necessarily fall along party lines (i.e. liberal outlets portraying liberal politicians positively), but these outlets do focus on different aspects of events.

Incident-Level For the incident-level analysis, we return to the example introduced in the beginning of this section: the accusations against comedian Aziz Ansari. First, we examine the power landscape surrounding Aziz Ansari through a graphical visualization. We develop this graph by drawing from psychology theories of power, which suggest that a person's power often derives from the people around them (French and Raven, 1959). We devise a metric to capture the concept of relative power: for any pair of entities, we average the difference between their power scores across all articles in which they are both mentioned. We visualize these differentials in a graph: an edge between two entities denotes that there is at least 1 article that mentions both entities. An edge from entity A to entity B denotes that on average, A is presented as more powerful than B. The magnitude of that difference is reflected in the edge weight. Finally, we sum all of the edge weights to obtain a score for each node. Thus, a large node for entity A indicates that entity A is often portrayed as more powerful than other entities mentioned in the same articles as A.

We show this graph for articles related to Aziz Ansari by limiting our corpus to articles that mention "Aziz" (555 articles). We take only the 100 most frequently mentioned entities, eliminate all non-people (i.e. Hollywood), and manually group redundant entities (i.e. Donald Trump vs. President Donald Trump). From this graph, we can see two separate clusters, related to two distinct narratives about Aziz Ansari. The cluster of entities in the top left corner, including Seth Meyers, Oprah Winfrey, and Nicole Kidman, is tied to media coverage of the 2018 Golden Globe Awards. All of theses entities won awards or gave speeches, including Aziz Ansari, who won Best Actor in a Television Series – Musical or Comedy. Much of the Golden Globes centered on the #MeToo

[13] https://dailycaller.com/2018/01/01/railroaded-the-real-reasons-al-franken-is-no-longer-a-senator/, https://freerepublic.com/focus/f-news/3611613/posts

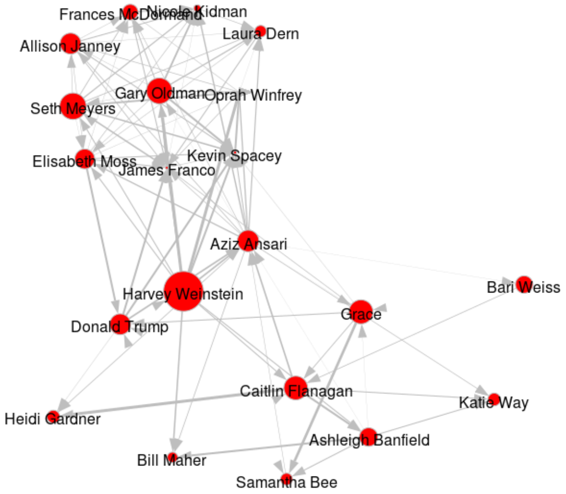


Figure 2.4: Graphical visualization of power dynamics between entities involved in the accusations against Aziz Ansari

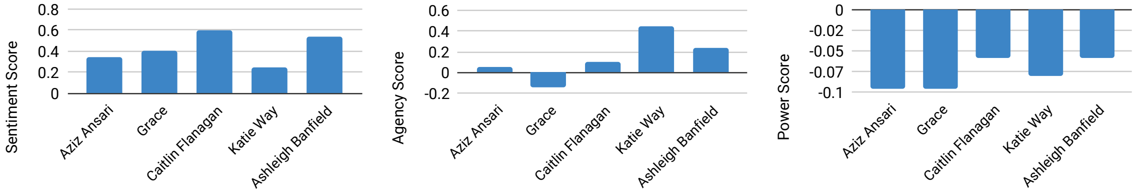


Figure 2.5: Power (left), agency (center), and sentiment (right) scores for common entities in articles about Aziz Ansari. Aziz Ansari and his accuser, Grace, are portrayed with comparable sentiment. Grace is portrayed with low agency, while journalist Katie Way has very high agency. Aziz Ansari and Grace are portrayed with lower power than journalist Ashleigh Banfield.

movement, as presenters expressed their support of the movement in speeches or wore pins indicating their support. Aziz Ansari himself wore a “Time’s Up” pin, to express his opposition to sexual harassment. Thus, because Aziz Ansari won a prominent award and much of the Golden Globes centered on the #MeToo movement, selecting all articles about Aziz Ansari from our #MeToo corpus results in many articles about the Golden Globes.

In the bottom right corner, we see a second narrative, relating to the accusations against Aziz Ansari. Aziz Ansari and Grace are the primary entities in the narrative. Other frequently mentioned entities include journalists: Katie Way, who authored the original article about the allegations, and Caitlin Flanagan and Ashleigh Banfield, who publicly criticized Katie Way’s article. We can see the importance of these journalists in their relative power. For instance, the edge from Caitlin Flanagan to Aziz Ansari indicates that she is often portrayed as more powerful than he is.

We then analyze the sentiment, power, and agency scores for these prominent entities, limiting our data set to articles that mention Aziz Ansari and were published after the Babe.net article that first disclosed the allegations against him (476 articles). Figure 2.5 shows that, unlike figures like Harvey Weinstein and Bill Cosby, who have much lower sentiment scores than their accusers, Aziz Ansari has a similar sentiment score to Grace, reflective of the mixed reaction to the article published

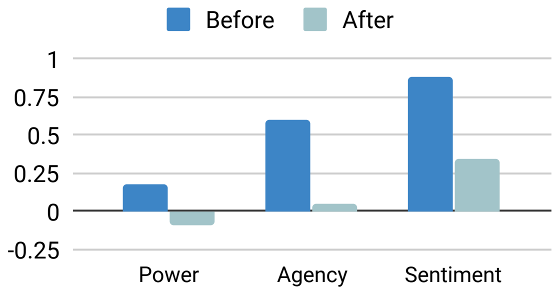


Figure 2.6: Aziz Ansari is portrayed with lower power, agency, and sentiment in articles published after the allegations against him.

about them. Katie Way, the journalist who wrote the original article, is portrayed with particularly low sentiment, which coincides with the severe criticism she received for publishing the story. In contrast, Caitlin Flanagan and Ashleigh Banfield, who were both front runners in criticizing Katie Way, were portrayed more positively.

Figure 2.5 also shows the agency scores for the same entities. Grace does have lower agency than Aziz Ansari, which supports the critique that Grace lacks any agency in the narrative. Aziz Ansari and Grace both have lower agency scores in comparison to all 3 journalists. Sentences where Ashleigh Banfield is scored as having high power and agency include: "Banfield slammed the accuser for her 'bad date'."[14] The high agency of these journalists demonstrates the prominent role that journalists have in social movements. In contrast Grace's low agency stems from sentences like "she felt pressured to engage in unwanted sexual acts." [15] The power scores (Figure 2.5, right) reflect a similar pattern, Grace and Aziz Ansari are portrayed as less powerful than the 3 journalists.

Finally, we compare sentiment, power, and agency scores for Aziz Ansari in articles published before the Babe.net article (79 articles) and articles published after (476 articles) in Figure 2.6. As articles about Aziz Ansari before these allegations focus primarily on his Golden Globe victory, these articles portray him with high power, agency, and sentiment. Following the Babe.net article, we see a decrease in Aziz Ansari's power, agency, and sentiment scores.

In this section, we introduce contextual affective analysis: a framework for analyzing nuanced entity portrayals. Our analysis of power, agency, and sentiment in media coverage of the #MeToo movement, addresses questions like "Whom does the media portray as sympathetic?" and "Whom does the media portray as powerful?" We demonstrate that although this movement has empowered women by encouraging them to share their stories, this empowerment does not necessarily translate into online media coverage of events. While women are among the most sympathetic entities, traditionally powerful men remain among the most powerful in media reports. We further show the prominence of journalists and 3rd party entities commenting on events without being directly involved, not only because their statements can influence perception of the movement, but because by making statements, they become entities in the narrative. Through this analysis, we highlight the importance of media framing: journalists can choose which narratives to highlight in order to promote certain portrayals of people. They can encourage or undermine movements like the #MeToo

[14] https://hollywoodlife.com/2018/01/16/wendy-williams-slams-aziz-ansari-accuser-sexual-misconduct-allegations-bad-date/

[15] https://www.hollywoodreporter.com/tv/tv-news/samantha-bee-fires-back-at-metoo-critics-addresses-aziz-ansari-accusations-1075595/

movement through their choice of entity portrayal.

2.2 Entity-centric Contextual Affective Analysis

2.2.1 Background

In §2.1, we develop metrics for scoring power, agency, and sentiment that are based on ternary verb scores. This approach is limited, in that it does not capture the implications of other parts of speech, like adjective and apposition nouns, which could also carry strong connotations. In this section, we leverage contextualized word representations to score entity mentions directly. We first describe our methodology (§2.2.2), which combines contextualized word embeddings with affect lexicons (Mohammad, 2018), and evaluate its performance, including discussing pros and cons of the entity-centric approach as compared to the verb-centric approach (§2.2.3). We then use our method to examine different portrayals of men and women (§2.2.4), focusing on the same domains as prior work (Wagner et al., 2015; Fu et al., 2016).

2.2.2 Methodology

Given an entity, such as "Batman", mentioned in a narrative, our goal is to obtain power, sentiment, and agency scores for the entity. We take two approaches: supervised regression and semi-supervised embedding projection. For both approaches, we use pre-trained contextualized embeddings as features and for training and test data we use the NRC Valence, Arousal, and Dominance (VAD) Lexicon, which contains valence (sentiment), arousal (agency), and dominance (power) annotations for more than 20,000 English words (Mohammad, 2018). It was created through manual annotations using Best–Worst scaling, and final annotations are on a scale from 0 (i.e. lower power) to 1 (i.e. high power). While we use this lexicon because its annotations contain our target dimensions of power, sentiment, and agency, our methodology readily generalizes to other lexicons.

Regression Model In the regression model, we take a supervised approach, using annotations from the NRC VAD Lexicon as training data.

Given a training word w and a large training corpus, we extract a contextual embedding $\mathbf{e}$ for every instance of w in the corpus. We use off-the-shelf pre-trained language models to extract sentence-level embeddings with no additional fine-tuning. Then, we average over all $\mathbf{e}$ embeddings for each instance w to obtain a single feature vector for each training point, and we train a Kernel Ridge Regression model using these embeddings as features.[16]

To extract affect scores for an entity in a narrative, we use the same pre-trained language model to extract a contextual embedding for the entity. Then, we feed this embedding through the regression model to obtain power, sentiment, and agency scores. When an entity occurs multiple times in the narrative, we average over the contextual embeddings for each occurrence of the entity and score the averaged embedding.

[16] We also experimented with Linear Regression and Ridge Regression, but found that Kernel Ridge Regression performed the best.

Power		Sentiment		Agency	
Low	High	Low	High	Low	High
timid	resourceful	negative	positive	silently	furiously
weakly	powerfully	pessimistic	optimistic	meek	lusty
cowardly	courageous	annoyed	amused	homely	sexy
inferior	superior	pessimism	optimism	bored	flustered
clumsy	skillful	disappointed	pleased	quietly	frantically

Table 2.7: Polar-opposite word pairs identified by ASP

Affect Subspace Projection (ASP) The main disadvantage of the regression approach is that we are unable to control for confounds and prevent overfitting to the training data. For example, many low-agency nouns tend to be inanimate objects (i.e. *table*), while high-agency nouns are people-oriented words (i.e. *dictator*). Thus, we can expect that the model learns to predict the difference between classes of nouns, rather than solely learning the affect dimension of interest. While other variations of regression allow for the inclusion of covariates and confounds, we have no systematic way to quantify or even identify these confounds. Instead, we devise a method to isolate dimensions of power, agency, and sentiment by first identifying corresponding subspaces in the embedding space and then projecting entities onto these dimensions. We refer to this method as *affect subspace projection* (ASP).

We describe this process for obtaining power scores; the agency and sentiment dimensions are analogous. In order to isolate the power subspace, we draw inspiration from Bolukbasi et al. (2016). First, we need to identify pairs of words whose meanings differ only in that one word connotes high power and the second word connotes low power. We define a set $\mathcal{H}$, which consists of the $|\mathcal{H}|$ highest-powered words from the VAD lexicon and a set $\mathcal{L}$, which consists of the $|\mathcal{L}|$ lowest powered words from the VAD Lexicon. For every word $w_h \in \mathcal{H}$, we use cosine similarity over contextual embedding representations to identify $w_l \in \mathcal{L}$, the low-powered word that is most similar to w_h. We allow each w_l to match to at most one w_h. Thus, we identify pairs of words (w_h, w_l), where w_h and w_l are very similar words but with polar opposite power scores. Finally, we keep only the N pairs with the greatest cosine similarity. We tune hyperparameters $|\mathcal{H}|$, $|\mathcal{L}|$, and N over a validation set. We show examples of extracted pairs for each dimension in Table 2.7.

Next, we use these paired words to construct a set of vectors whose direction of greatest variance is along the power subspace. For each pair of high and low power words (w_h, w_l), we take their embedding representations $\mathbf{e_h}$ and $\mathbf{e_l}$ in the same way as in the regression model. We then define $\mu = (\mathbf{e_h} + \mathbf{e_l})/2$, and construct a matrix $\mathbf{M}$, where each row is $\mathbf{e_l} - \mu$ or $\mathbf{e_h} - \mu$. Thus, $\mathbf{M}$ is a $d \times 2N$ dimensional matrix, where d is the dimension of the embeddings. We then run PCA over $\mathbf{M}$ to extract its principle components and keep the first principle component as the target subspace.

Finally, to score an entity in a narrative, we take the entity's contextual embedding representation and project it onto the identified subspace. Because we keep only the first principle component as the target subspace, the projection results in a single-dimensional vector, i.e., a power score. We repeat the process for agency and sentiment, constructing 3 separate $\mathbf{M}$ matrices in order to obtain power, sentiment, and agency scores.

	Regression			ASP		
	Power	Sentiment	Agency	Power	Sentiment	Agency
ELMo	0.78	**0.84**	0.76	0.65	0.76	0.63
BERT	**0.79**	0.83	**0.78**	0.65	0.71	0.66
BERT-masked	0.64	0.70	0.62	0.41	0.47	0.41

Table 2.8: Pearson correlations between gold NRC VAD labels and scores predicted by our models. Correlations are generally high, with the regression method outperforming ASP. All correlations are statistically significant ($p < 1e - 75$).

2.2.3 Evaluation

As in §2.1.3 we first evaluate our method's ability to re-create held-out lexicon data, and then we evaluate its ability to capture entity portrayals. In Field and Tsvetkov (2019), we additionally provide a more qualitative evaluation, which is omitted here for brevity. To train and evaluate our methods, we randomly divide the VAD lexicon into train (16,007), development (2,000), and test (2,000) sets. We extract embeddings to train our models from a corpus of 42,306 Wikipedia movie plot summaries (Bamman et al., 2013). When experimenting with other training corpora, such as newspaper articles, we found the choice of training corpus had little impact on results. We use two pretrained language models to extract embeddings: ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). It is important to note that the movie plots corpus we used for extraction is not identical to the corpora used to train ELMo (5.5B tokens from Wikipedia and WMT news crawl) and BERT (800M-word BooksCorpus and 2,500M-word Wikipedia).

We use two variants of BERT to extract embeddings. In the first, referred to as "BERT-masked", we mask out the target word before extracting embeddings from an input sentence. Masking out target words is a part of the BERT training objective (Devlin et al., 2019). By using masks in our embedding extractions, we force the model to produce an embedding solely from the context surrounding the word, rather than relying on information from the word itself. In the second variant, referred to as "BERT", we extract embeddings over each sentence containing a target without modification. Field and Tsvetkov (2019) further details including hyperparamter settings.

Lexicon Evaluation Table 2.8 shows the Pearson correlations between gold annotations and the scores predicted by our models over the held-out VAD test set. The high correlations demonstrate that both the regression and ASP models successfully capture information about power, sentiment, and agency from contextualized embeddings. The ELMo embeddings and unmasked BERT embeddings perform approximately the same. However, the masked BERT embeddings perform markedly worse than the unmasked embeddings.[17] The poorer performance of the masked embeddings demonstrates the extent to which the BERT model biases representations towards the actual observed word, which is explicitly one of the motivations of the BERT training objective Devlin et al. (2019). More specifically, when we mask out the target before extracting embeddings, we force the extracted embedding to only encode information from the surrounding context. Then any improvements in performance when we do not mask out the target are presumably obtained from the word-form for

[17] One of the drawbacks of context-based word embeddings is that antonyms like "positive" and "negative" tend to have similar embeddings, because they tend to be used in similar contexts. However, given the breadth of words in the VAD lexicon, we do expect context to differ for oppositely scored words. For instance we would expect "pauper" and "king" to be used in different contexts, as well as "pauper" and "powerful".

	Regression	ASP
ELMo	0.51	0.21
BERT	0.38	0.38
BERT-masked	0.17	-0.085
ELMo + Freq	**0.65**	0.48
Frequency Baseline	0.61	
Verb-Centric approach (§2.1.2)	-0.12	

Table 2.9: Spearman correlations between automatically induced power scores and Forbes power ranking. Correlations for ELMo regression ($p = 0.029$), ELMo regression + Freq ($p = 0.003$), and the frequency baseline ($p = 0.007$) are statistically significant. The ELMo regression + Freq model performs the best.

the target itself. For example, we may score "king" as high-powered because "king" often occurred as a high-powered entity in the data used to train the BERT model, regardless of whether or not it appeared to be high-powered in the corpus we ultimately extract embeddings from. Nevertheless, training with BERT-masked embeddings still results in statistically significant correlations, which suggests that some affect information is derived from surrounding context.

The regression model generally outperforms ASP on this task. The regression model has an advantage over ASP in that it is directly trained over the full lexicon, whereas ASP chooses a subset of extreme words to guide the model. However, as discussed in §2.2.2, it is difficult to determine what effect other confounds have on the regression model, while the ASP approach provides more concrete evidence that these contextualized word embeddings encode affect information.

In-domain Entity Evaluation Next, we evaluate how well our models capture affect information in entities, rather than words, by assessing power scores through two metrics. We compare our models against the verb-centric approach presented in §2.1.2 and against a frequency baseline, where we consider an entity's power score to be the number of times the entity is mentioned in the text.

First, we consider an in-domain task, where we compare our metrics for scoring power with a standard benchmark that we expect to be reflected in both the data we use to extract embeddings and the data used to train ELMo and BERT. More specifically, we use the power scores obtained from our model to rank the 20 most powerful people in 2016 according to Forbes Magazine.[18]

This is a particularly difficult task: unlike in §2.1.3 where we seek to identify pairwise power relations in a set of articles, we seek to directly rank people according to their power, which requires more precise scores. Furthermore, the frequency metric supplies a particularly strong baseline. The metrics that Forbes Magazine uses to compose the list of powerful people include a person's influence as well as how actively they use their power.[19] Under these conditions, Forbes Magazine may consider a person to be powerful simply because they are mentioned frequently in the media. Additionally, we can surmise that people who actively use their power are mentioned frequently in the media.

Table 2.9 presents Spearman correlations between our scores and rank on the Forbes list for each model. For all metrics, we construct embeddings from every instance of each person's full

[18] https://www.forbes.com/sites/davidewalt/2016/12/14/the-worlds-most-powerful-people-2016/?sh=664ddb9c1b4c

[19] https://www.forbes.com/sites/davidewalt/2018/05/08/the-worlds-most-powerful-people-2018/?sh=556e1f016c47

	Full set (383 pairs)		Reduced set (49 pairs)	
	Regression	ASP	Regression	ASP
ELMo	44.9	43.6	36.7	42.8
BERT	41.8	49.3	42.9	49.0
BERT-masked	49.6	**59.0**	53.1	55.1
Frequency Baseline	58.0		57.1	
Verb-centric approach	-		**71.4**	

Table 2.10: Accuracy for scoring how powerful entities are as compared with annotations over articles related to the #MeToo movement. Our metrics do not consistently outperform the baselines, suggesting ELMo and BERT embeddings fail to transfer across domains.

name in U.S. articles from 2016 in the NOW news corpus.[20] In addition to the proposed methods, we used our best performing model (regression with ELMo) to augment the frequency baseline, by normalizing and summing the frequency scores with the scores from this model. This combined model achieves the strongest correlation (raw scores from this model are shown in Figure 2.7). Furthermore, the regression with ELMo model alone achieves a statistically significant correlation even without the incorporation of frequency scores. The unmasked BERT embeddings also achieve positively correlated scores, though these correlations are not statistically significant. The BERT-masked embeddings perform particularly poorly, as does the verb-centric approach. While the verb-centric approach may be capable of identifying powerful entities, we suspect it is not fine-grained enough to rank them.

Frequency serves as a strong baseline for power, but we would not expect frequency to be a good measure of sentiment or agency. None of our metrics for these traits are significantly correlated with the Forbes' ranking. Also, we would not expect frequency to be a good measure in other contexts, such as how powerfully an entity is portrayed in a single document rather than across a large media corpus.

Out-of-domain Entity Evaluation Next, we entity-centric approach using the same evaluation task introduced in §2.1.3. We compare our scores with hand-annotated power rankings over a set of newspaper articles related to allegations of sexual harassment against the comedian Aziz Ansari. This setting constitues an out-of-domain evaluation task, as we do not expect people portrayals in this specific incident to align with portrayals in the data used to train ELMo and BERT.

The verb-centric approach restricts analysis to a limited set of pairs, since only entities used with verbs from the lexicon are included. In contrast, in the entity-centric approach, we simply use string matching to identify entities in the text, without requiring that the entities be linked to specific verbs, allowing for the identification of more entities. Table 2.10 shows results over the same set of pairs used for evaluation in Table 2.4 as well as an expanded set, when we do not restrict to entities used with lexicon verbs. The entity-centric approaches fail to consistently outperform even the frequency baseline for this task, likely because the ELMo and BERT embeddings are biased towards their training data.

The #MeToo movement is widely known for subverting traditional power roles: allegations made by traditionally unpowerful women have brought down traditionally powerful men. For example, Harvey Weinstein, an influential film producer, has traditionally been a powerful figure in society, but

[20] https://corpus.byu.edu/now/

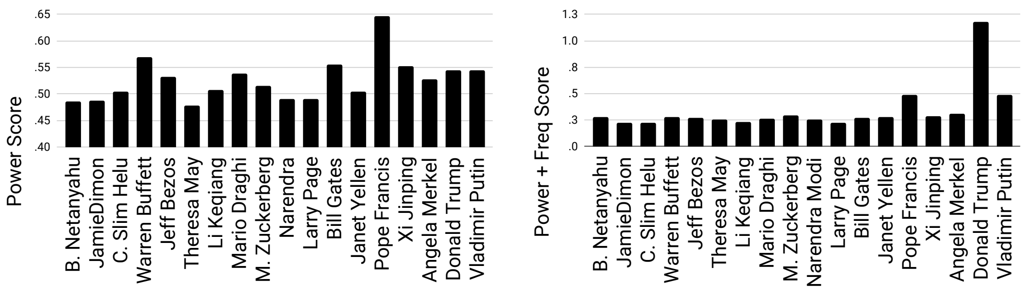


Figure 2.7: Power scores for people on the 2016 Forbes Magazine power list (omitting two people who occur infrequently in the corpus) as learned through regression with ELMo embeddings, and through combined regression and frequency scores. Women are generally scored lower than similarly ranked men.

numerous allegations of sexual harassment have resulted in his effective removal from the industry. We can speculate that the broad corpora used to train ELMo and BERT portray these men as more powerful than articles about the #MeToo movement.

Thus, in a corpus where there may be differences in traditional power roles, the embeddings extracted from ELMo and BERT perform worse than random, as they are biased towards the power structures in the data they are trained on. Further evidence of this exists in the performance of the BERT-masked embeddings - whereas these embeddings generally capture power poorly as compared to the unmasked embeddings (Table 2.8), they outperform the unmasked embeddings on this task, and even outperform the frequency baseline in one setting. Nevertheless, they do not outperform the verb-centric approach, likely because they do not capture affect information as well as unmasked embeddings.

2.2.4 Analysis of Gender Bias in Media

In §2.2.3 we discuss some of the trade-offs of the verb-centric approach and entity-centric approaches, and suggest that the verb-centric approach is more useful in examining connotations in specific corpora, but that the entity-centric may better capture connotations in broad corpora that align with ELMo and BERT training data. In this section, we provide an example of the type of analyses suited to the entity-centric approach by examining how men and women are portrayed in the media, focusing on domains of interest in prior NLP work (Wagner et al., 2015; Fu et al., 2016). We use the NOW corpus and regression with ELMo embeddings for analysis.[21]

First, we return to the example from Table 2.9, the list of most powerful people from Forbes Magazine in 2016. Figure 2.7 shows the power scores ordered from least powerful to most powerful according to the Forbes list. We show both the raw power scores computed by our model, as well as the regression power scores combined with frequency metric (as in Table 2.9). In the raw scores, stand-out powerful people include businessman Warren Buffet and Pope Francis. In contrast, the only 3 women, Theresa May, Janet Yellen, and Angela Merkel, are underscored as compared to similarly ranked men. However, when we incorporate frequency, we do not see the same underscoring. This result suggests that although these women are portrayed frequently in the media, they are

[21] ASP results are nearly identical.

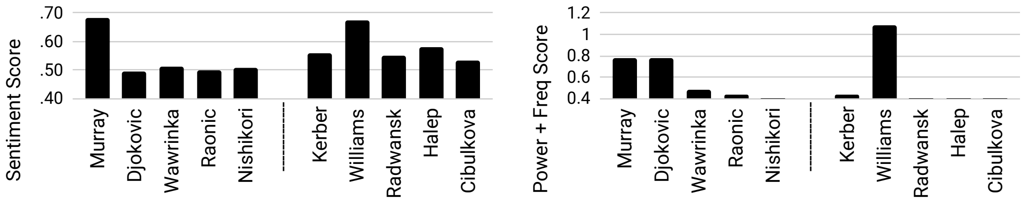


Figure 2.8: Sentiment (left) and power scores (right) for the top-ranked male (left) and female (right) tennis players in 2016 through regression with ELMo embeddings (power scores combine regression scores with frequency counts). Women are generally portrayed with lower power and higher sentiment.

typically described as less powerful than their actual role in society.[22] This finding is consistent with prior work on portrayals of women (Wagner et al., 2015). The most striking difference after the incorporation of frequency scores is the boosted power score for Donald Trump, who is mentioned much more frequently than other entities.

In Figure 2.8, we show the sentiment and power (combined regression + frequency) scores for the top-ranked male and female tennis players in 2016. Prior work has shown bias in news coverage of male and female tennis players, specifically, that male players are typically asked questions more focused on the game than female players (Fu et al., 2016). Our analysis focuses on a different data set and coverage type—we examine general articles rather than post-match interviews. As expected, popular players Serena Williams and Andy Murray have the highest sentiment scores and very high power scores. In contrast, Novak Djokovic, who has notoriously been less popular than his peers, has the lowest sentiment score, but the second highest power score (after Williams). Additionally, female players are typically portrayed with more positive sentiment (female average score = 0.58; male average score = 0.54), whereas male players are portrayed with higher power (female average score = 0.52; male average score = 0.57). However, the difference in power disappears when we remove frequency from the metric and use only the regression scores, suggesting that the difference occurs because male players are mentioned more frequently.

In this section, we propose a method for incorporating contextualized word embeddings into entity-centric analyses, which has direct applications to social analyses. Our results are easy to interpret and readily generalize to a variety of research questions. However, we further expose several limitations to this method, specifically that contextualized word embeddings are biased towards representations from their training data, which limits their usefulness in new domains. Despite this limitation, we find that these models are expressive enough to analyze entity portrayals in in-domain data, allowing us to examine different portrayals of specific men and women.

[22] We note that the portrayals of other people with the same first names in the training data may have biased ELMo embeddings

2.3 Controlling for Confounding Variables in Analyses of People

2.3.1 Background

§2.1 and §2.2 provide methodology for analyzing how people are portrayed in narrative next and show how these methods can be used to examine specific individuals in large corpora. However, in examining signs of prejudice and social bias, we are often interested in broad group-level analyses (e.g. are men or women portrayed as more powerful?). However, when we make comparisons between large groups of people, attributes about people other than their gender limit conclusions. For example, much prior work has examined gender bias in Wikipedia articles, with findings including: articles about women tend to be longer than articles about men (Graells-Garrido et al., 2015; Reagle and Rhue, 2011; Wagner et al., 2015; Young et al., 2016), all biography articles tend to link to articles about men more than women (Young et al., 2016; Wagner et al., 2015, 2016; Eom et al., 2015), and pages for women discuss personal relationships more frequently than pages for men (Bamman and Smith, 2014; Graells-Garrido et al., 2015; Wagner et al., 2016). However, there are more male than female athletes on Wikipedia, so it is difficult to disentangle if differences occur because women and men are presented differently, or because non-athletes and athletes are presented differently (Graells-Garrido et al., 2015; Wagner et al., 2016; Hollink et al., 2018). This entanglement is evident when examining word statistics, which show that over-represented words on pages for men as opposed to women are sports terms: "footballer", "baseball", "league" (Graells-Garrido et al., 2015; Wagner et al., 2016).

Here, we develop a *matching algorithm* that enables analyses by isolating target dimensions. Given a set of people with a *target* attributes (e.g. cisgender women), our algorithm builds a comparison set of people that do not (e.g. cisgender men). We construct this comparison set to closely match the target set on all known attributes except the targeted one (e.g. gender) by using pivot-slope TF-IDF weighted attribute vectors. Thus, examining differences between the two corpora can reveal *content bias* (Young et al., 2016) related to the target attribute.

In this section, we focus on isolating content bias in Wikipedia biography pages. Wikipedia is a widely-read global platform and has become a popular data source of NLP training data. Thus, biases in Wikipedia articles can influence readers and also be absorbed and amplified by computational models (Bolukbasi et al., 2016; Zhao et al., 2017; Peters et al., 2018; Mora-Cantallops et al., 2019; Redi et al., 2020). Concerns about social biases on Wikipedia have lead to much prior research on this topic, primarily focused on binary gender (Reagle and Rhue, 2011; Bamman and Smith, 2014; Graells-Garrido et al., 2015; Wagner et al., 2015, 2016; Young et al., 2016; Hollink et al., 2018). Some of this prior work has drawn the attention of the editor community and led to changes on the platform (Reagle and Rhue, 2011; Langrock and González-Bailón, 2020), which further motivates our focus on this platform. However, the methods we present are applicable to any other corpora with sparse high-dimensional confounding variables.

We first present our matching methodology (§2.3.2) and evaluate over a large Wikipedia biography corpus, focusing on isolating race and gender bias (§2.3.3). We then present an example analysis of how this matching approach can be combined with context affective analysis to examine how prejudice about LGBT people manifests on Wikipedia (§2.3.4).

2.3.2 Methodology

Matching Methodology We present a method for identifying a *comparison* biography page for every page that aligns with a target attribute, where the comparison page closely matches the target page on all known attributes except the target one. The concept originates in adjusting observational data to replicate the conditions of a randomized trial; from the observational data, researchers construct treatment and control groups so that the distribution of all covariates except the target one is as identical as possible between the two groups (Rosenbaum and Rubin, 1983).[23] Then by comparing the constructed treatment and control groups, researchers can isolate the effects of the target attribute from other confounding variables. Matching is also gaining attention in language analysis (Choudhury et al., 2016; Chandrasekharan et al., 2017; Egami et al., 2018; Roberts et al., 2020; Keith et al., 2020). Here, our target attribute is gender or race as likely to be perceived by editors and readers. We aim to create corpora that balance other characteristics, such as age, occupation, and nationality, that could affect how articles are written.

Given a set of target articles $\mathcal{T}$ (e.g. all biographies about women), our goal is to construct a set of comparison articles $\mathcal{C}$ from a set of candidates $\mathcal{A}$ (e.g. all biographies about men), such that $\mathcal{C}$ has a similar covariate distribution as $\mathcal{T}$ for all covariates except the target attribute. We construct $\mathcal{C}$ using greedy matching. For each $t \in \mathcal{T}$, we identify $c_{best} \in \mathcal{A}$ that best matches t and add c_{best} to $\mathcal{C}$. If t is about an American female actress, c_{best} may be about an American male actor. To identify c_{best}, we leverage the category metadata associated with each article. For example, the page for Steve Jobs includes the categories "Pixar people", "Directors of Apple Inc.", "American people of German descent", etc. While articles are not always categorized correctly or with equal detail, using this metadata allows us to focus on covariates that are likely to reflect the way the article is written. People may have relevant traits that are not listed on their Wikipedia page, but if no editor assigned a category related to the traits, we have no reason to believe editors were aware of them nor that they influenced edits. We describe 6 methods for identifying $c_{best} \in A$. $CAT(c)$ denotes the set of categories associated with c.

<u>Number</u> We choose c_{best} as the article with the most number of categories in common with t, which is intuitively the best match.

<u>Percent</u> Number favors articles with more categories. For example, a candidate c_i that has 30 categories is more likely to have more categories in common with t than a candidate c_j that only has 10 categories. However, this does not necessarily mean that c_i has more traits in common with the person t—it suggests that the article is better written. We can reduce this favoritism by normalizing the number of common categories by the total number of categories in the candidate c_i, i.e. $c_{best} = \arg\max_{c_i} |CAT(c_i) \cap CAT(t)| \frac{1}{|CAT(c_i)|}$

<u>TF-IDF</u> Both prior methods assume that all categories are equally meaningful, but this is an oversimplification. A candidate c_i that has "American short story writers" in common with t is more likely to be a good match than one with "Living People" in common. We use TF-IDF weighting to up-weight rarer categories (Salton and Buckley, 1988). We represent each $c_i \in \mathcal{A}$ as a sparse category vector, where each element is a product between the frequency of the category in c_i, ($\frac{1}{|CAT(c_i)|}$ if the category is in $CAT(c_i)$, 0 otherwise) and the inverse frequency of the category, which down-weights broad common categories. We select c_{best} as the c_i with the highest cosine similarity between its

[23] We use target/comparison instead of treatment/control to clarify that our work does not involve any actual "treatment".

vector and a similarly constructed vector for t.

PROPENSITY For each article we construct a propensity score, an estimate of the probability that the article contains the target attribute (Rosenbaum and Rubin, 1983, 1985), using a logistic regression classifier trained on one-hot-encoded category features. We then choose c_{best} as the article the closest propensity score to t's. Propensity matching is not ideal in our setting, first, because it was designed for lower-dimensional covariates and has been shown to fail with high-dimensional data, and second, because it does not necessarily produce matched pairs that are meaningful, which precludes manually examining matches (Roberts et al., 2020). Nevertheless, we include it as a baseline, because it is a popular method for controlling for confounding variables.

TF-IDF PROPENSITY We construct an additional propensity score model, where we use TF-IDF weighted category vectors as features instead of one-hot encoded vectors.

PIVOT-SLOPE TF-IDF TF-IDF and PERCENT both include the term $\frac{1}{|CAT(c_i)|}$ to normalize for articles having different numbers of categories. However, information retrieval research suggests that it over-corrects and causes the algorithm to favor articles with fewer categories (Singhal et al., 1996). Instead, we adopt pivot-slope normalization (Singhal et al., 1996) and normalize TF-IDF terms with an adjusted value: $(1.0 - slope) * pivot + slope * |CAT(c_i)|$. This approach requires setting the slope and the pivot, which control the strength of the adjustment. Following Singhal et al. (1996), we set the pivot to the average number of categories across all articles, and tune the slope over a development set. Tuning the slope is important, as changing the parameter does change the selected matches. PIVOT-SLOPE TF-IDF is our novel proposed approach.

In practice, it is likely not possible to identify close matches for every target article, i.e. there may be characteristics of people in the target corpus that are not shared by anyone in the comparison corpus. To account for this, we discard "weak matches": for direct matching methods, pairs with < 2 categories in common, for propensity matching methods, pairs whose difference in propensity scores is > 1 standard deviation away from the mean difference. Field et al. (2022) provides details on experimental setups.

Model Assumptions and Limitations In this section, we clarify how some of the assumptions required by our methodology limit the way it should be used. First, our matching method depends on categories, which are imperfect controls. While we take some steps to account for this, e.g., excluding articles with few categories, systemic differences in category tagging could reduce the reliability of matching (we did not observe evidence of this). Using categories as covariates also precludes us from identifying systemic differences in how article categories are assigned. Instead our work focuses on differences in articles, given the current category assignments. Second, our methodology is only meaningful where it is possible to establish high-quality matches. If there are people with characteristics in our target corpus that do not present in our comparison corpus, we cannot draw controlled comparisons. In practice, we operationalize this limitation by discarding pairs of articles that are poorly matched (described in §2.3.2) and only computing analyses over sets of articles where there is covariate overlap. Third, we note that while we borrow some methodology and terminology from causal inference, our setup is not conducive to a strictly causal framework, and we do not suggest that all results imply causal relations. It is difficult to determine if article imbalances are the result of Wikipedia editing, societal biases, or other factors, meaning there are confounding variables we cannot control for. To summarize, we aim to identify systemic differences in articles about different groups of people, and we view the main use case of our model as identifying sets of articles that may benefit from manual investigation and editing.

Evaluation Framework We devise schemes to evaluate both how well each matching metric creates comparison groups with similar attribute distributions as target groups and if the metric introduces imbalances, e.g. by favoring articles with fewer categories. Given matched target and comparison sets, we assess matches using several metrics:

Standardized Mean Difference (SMD) SMD (the difference in means between the treatment and control groups divided by the pooled standard deviation) is the standard method used to evaluate covariate balance after matching (Zhang et al., 2019). We treat each category as a binary covariate that can be present or absent for each article. We then compute SMD for each category and report the average across all categories (AVG SMD) as well as the percent of categories where SMD>0.1. There is no widely accepted standardized bias criterion, but prior work suggests 0.25 or 0.1 as reasonable cut-offs (Harder et al., 2010). High values indicate that the distribution of categories is very different between the target and the comparison groups.

Number of Categories As discussed in the preceding paragraphs, the proposed methods may favor articles with more or fewer categories. Thus, we compute the SMD between the number of categories in the target group and comparison groups. A high value indicates favoritism.

Text Length The prior two metrics focus on the categories, but categories are a proxy for confounds in the text. We ultimately seek to assess how well matching controls for differences in the actual articles. We first compare article lengths (word counts) using SMD.

Polar Log-odds (PLO) We use log-odds with a Dirichlet prior (Monroe et al., 2008) to compare vocabulary differences between articles, where high log-odds polarities indicate dissimilar vocabulary.

KL Divergence Beyond word-level differences, we compute the KL-divergence between 100-dimensional topic vectors derived from an LDA model (Blei et al., 2003) in both the target-comparison (KL) and the comparison-target directions (KL 2).

We compute these metrics over three types of target sets:

Article-Sampling We construct simulated target sets by randomly sampling 1000 articles. Because we do not fix a target attribute, we expect a high quality matching algorithm to identify a comparison set that matches very closely to the target set, without creating imbalances, e.g. by favoring longer articles with more categories.

Category-Sampling We randomly sample one category that has at least 500 members, and then sample 500 articles from the category. We do not expect there to be any bias towards a single category, since most categories are very specific, e.g. “Players of American football from Pennsylvania”. While articles for football players might have different characteristics than other articles, we would not expect articles for players from Pennsylvania to be substantially different than articles for players from New York or New Jersey. Thus, as in the article-sampling setup, we can evaluate both attribute distributions and artificial imbalances. However, this setup more closely replicates the intended analysis, as we ensure that all people in the target group have a common trait.

Attribute-specific We evaluate how well each method balances covariates in analysis settings, e.g., when comparing articles about women and men. In this setting, we only consider how well the method balances covariates (SMD), using heuristics to exclude categories that we expect to differ between groups (e.g., when comparing cis. men and women, we exclude categories containing

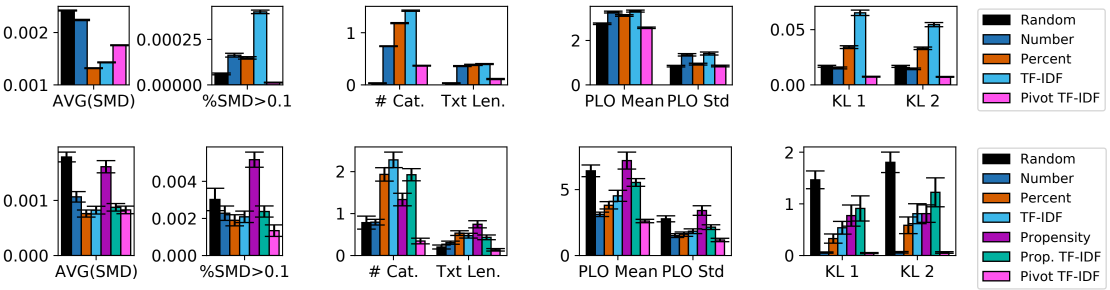


Figure 2.9: Evaluation of matching methods using article-sampling (top) and category-sampling (bottom), with 99% confidence intervals computed over 100 simulations. Lower scores indicate better matches; pivot-slope TF-IDF performs best overall. We omit propensity matching in article-sampling, as it is not meaningful.

Group	Pre-match	Final
African American	9,668	8,404
Asian American	4,728	3,473
Hispanic/Latinx American	4,483	3,813
Unmarked American (comparison)	93,486	-
Non-Binary	200	127
Cisgender women	108,915	64,828
Transgender women	261	134
Transgender men	85	53
Cisgender men (comparison)	331,484	-

Table 2.11: Data set sizes for analysis corpora. “Final” column indicates the target/comparison sizes after corpora are matched using Pivot-Slope TD-IDF and matches with < 2 categories in common are discarded.

the word “women”). We cannot examine other criterion, such as text length, because we cannot distinguish if differences between the target and comparison sets are signs of social bias or poor matching, especially considering prior work has suggested text length differs for people of different genders (Graells-Garrido et al., 2015; Reagle and Rhue, 2011; Wagner et al., 2015; Young et al., 2016). Instead, we use the synthetically constructed *article-sampling* and *category-sampling* to examine signs of favoritism in the algorithm and how well the matching controls for confounds in the article text.

2.3.3 Evaluation

Data In order to facilitate evaluation and analysis, we built a corpus of Wikipedia biography articles. We gathered all articles with the category “Living people” on English Wikipedia in March 2020. We discarded articles with < 2 categories, < 100 tokens, or marked as stubs (containing a category like “Actor stubs”). We use English categories for matching, which we expect to be the most reliable, because English has the most active editor community. We ignore categories focused on article properties instead of people traits using a heuristics, e.g., categories containing “Pages with”. Our final corpus consists of 444,045 articles, containing 9.3 categories and 628.2 tokens on average. The total number of categories considered valid for matching is 209,613.

We additionally infer race and gender of article subjects. Identity traits are fluid, difficult to

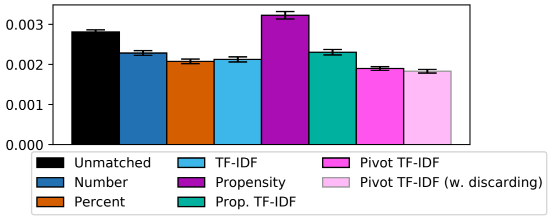


Figure 2.10: SMD between articles about African American people and matched comparisons, averaged across categories with 99% confidence intervals. Lower scores indicate a better match. "(w. discarding)" indicates SMD after weak matches are discarded.

operationalize, and depend on social context (Bucholtz and Hall, 2005; Hanna et al., 2020). Our goal is to identify *observed* gender and race as perceived by Wikipedia editors who assigned article metadata or readers who may view them, as opposed to assuming ground-truth values (Roth, 2016). Thus, we derive race and gender directly from Wikipedia, using a combination of article categories and other associated metadata (Vrandečić and Krötzsch, 2014). We treat identity as fixed at the time of data collection and caution that identity traits inferred in this work may not be correct in other time-periods or contexts. Field et al. (2022) provide details on this process, and Table 2.11 provide final data set sizes.

Article-sampling and Category-sampling Evaluation Figure 2.9 reports evaluation results for 100 *article-sampling* and *category-sampling* simulations. In addition to the described matching algorithms, we show the results of randomly sampling a comparison group. All evaluation metrics measure differences between the target and comparison groups: lower values indicate a better match. Throughout all evaluations, except where explicitly noted, we do not exclude weak matches in order to retain comparable target sets. Exclusion can result in different target articles being discarded under different matching approaches.

All methods perform better than random in reducing covariate imbalance, and Pivot-Slope TF-IDF best reduces the percentage of highly imbalanced categories (%SMD>0.1). In the category-sampling simulations (bottom), which better simulate having a target group with a particular trait in common, all methods also perform better than random in the text-based metrics (PLO and KL), and Pivot-Slope TF-IDF performs best overall. In article-sampling simulations (top), random provides a strong baseline. This is unsurprising, as randomly chosen groups of 1000 articles are unlikely to differ greatly. Nevertheless, Pivot-Slope TF-IDF outperforms random on covariate balancing and the text-based metrics.

As expected, Number exhibits bias towards articles with more categories, while Percent and TF-IDF exhibit bias towards articles with fewer categories, resulting in worse performance than random over the the number of categories (# Cat.) metric (Figure 2.9 reports absolute values). These differences are also reflected in text length, as articles with more categories also tend to be longer. In category-sampling, pivot-slope normalization corrects for this length bias and outperforms random. In article-sampling, while Pivot-Slope TF-IDF does outperform other metrics, the random sampling exhibits the least category number and text length bias. However, as mentioned, random is a strong baseline in this setting.

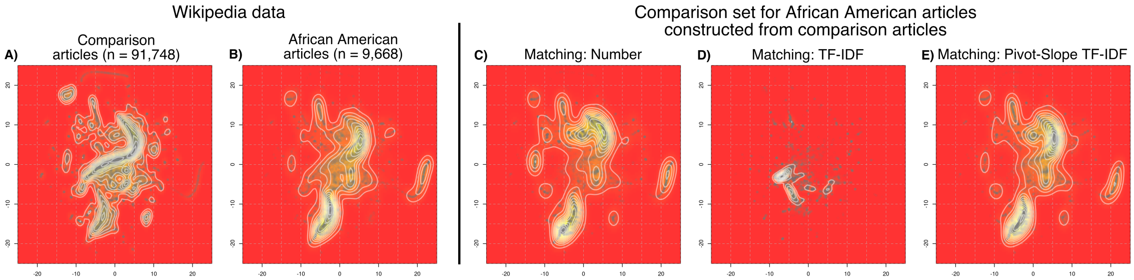


Figure 2.11: UMAP visualizations of TF-IDF weighted categories, where each point represents an article, and the red-to-yellow coloring depicts the low-to-high density of articles emphasized by white contours. (A) and (B) shows category distributions in African American and (unmatched) comparison articles respectively. (C) - (E) show category distributions in comparison sets generated by different matching methods. Percent and propensity methods (not shown for brevity) perform strictly worse. Pivot-Slope TF-IDF (D) results in the comparison set with the most similar category distribution as the target set (B).

Attribute-specific Evaluation In Figure 2.10 presents SMD averaged across categories between articles about African American people and comparisons, under different matching methods (evaluations for other target groups in Field et al. (2022) show similar patterns). As in Figure 2.9, we generally do not exclude weak matches in order to have directly comparable target sets, though we do report SMD after this exclusion for Pivot-Slope TF-IDF in order to accurately reflect what the final SMD would be. We further note that if we discard weak matches for all methods, Pivot-Slope TF-IDF results in the least amount of discarded data. With the exception of Propensity, all methods improve covariate balance as compared to not using matching, and Pivot-Slope TF-IDF performs best.

To supplement the quantitative results, we additionally provide qualitative results via UMAP (McInnes et al., 2018). Given a matrix $X \in \mathcal{R}^{n*k}$ with n rows (i.e., articles) and k columns (i.e., categories), UMAP nonlinearly maps each row into a two-dimensional space that preserves nearest-neighbor geometry. UMAP is often preferred over other nonlinear dimension reduction methods for its effectiveness in visualizing the data and assessing clustering structure, e.g., in text-analyses (Ochigame and Ye, 2021; Bolukbasi et al., 2021) and genomics (Becht et al., 2019). We set X to be the matrix of TF-IDF weighted categories for all methods to obtain a global set of n coordinates (one per article).

Figure 2.11 shows UMAP visualizations for African Americans. Without matching, the set of all comparison articles (A) has a distinctly different distribution of associated categories than articles about African American people (B), which motivates our work. Of the matching methods, Pivot-Slope TF-IDF (E) produces a comparison set with the most similar distribution of associated categories as those from articles of African American people. While Number (C) does also produce a similar category distribution, Figure 2.9 demonstrates that this method favors articles with more categories. TF-IDF (D) performs particularly poorly. As this method overly-favors articles with too few categories, comparison articles with few categories appear in a disproportionate number of matches.

Evaluating Effects on Analysis Finally, we consider how analysis results can differ without matching. Table 2.12 revisits the example in §2.3.1 and presents the words most associated with biography pages about cisgender men and women calculated using log-odds (Monroe et al., 2008). As

Unmatched (M)	he/He, his/His, **season**, him, **League**, **club**
Unmatched (W)	her/Her, she/She, women, actress, husband
Matched (M)	he/He, his/His, him, himself, wife, Men
Matched (W)	her/Her, she/She, women, husband, female

Table 2.12: Log-odds scores between cis. men and women pages ordered most-to-least polar from left to right. Matching reduces sports terms (bold) in favor of overtly gendered terms.

	Without Matching		Pivot-Slope TD-IDF Matching	
	Target	Comparison	Target	Comparison
African American	902.0	**711.4**	942.9	959.2
Asian American	737.5	**711.4**	**792.3**	854.1
Hispanic/Latinx American	972.5	**711.4**	3,813	1017.2

Table 2.13: Average article lengths with and without matching. Without matching, all target sets appear significantly longer than comparisons. With matching, articles about Asian American people are significantly shorter than comparisons. When differences are significant, the smaller value is in bold (p<0.05).

shown in previous work, without matching, words highly associated with men include many sports terms, which suggests that directly comparing these biographies could capture athlete/non-athlete differences rather than man/woman differences. After matching, these sports terms are replaced by overtly-gendered terms like “himself” and “wife”, showing that matching helps isolate gender as the primary variable of difference between the two corpora. Beyond the top log-odds scores, sports terms do occur, but they tend to be more specific and represented on both sides, for example “WTA” is women-associated and “NBA” is men-associated.

Table 2.13 shows the results of comparing article lengths between racial subgroups and comparison articles. Without matching, articles for all racial subgroups are significantly longer than comparison articles. However, after matching, we do not find significant differences between matched comparison articles and articles about African American and Hispanic/Latinx American people. Instead, we find that articles about Asian American people are typically shorter than comparison articles.

2.3.4 Analysis of Multilingual Wikipedia portrayals of LGBT People

Finally, we demonstrate how pivot-slope TF-IDF matching can be combined with contextual affective analysis in order to examine prejudice in text corpora. Specifically, we examine portrayals of LGBT people on Wikipedia. We focus on LGBT people because discrimination against the LGBT community is an increasingly important global issue. Although pride marches are held ≥ 158 cities world-wide Lisitza (2017), social oppression is prevalent in many countries (Balsam et al., 2011; Doi and Stewart, 2019). Nevertheless, little prior computational work has studied narratives about the community, likely due to data scarcity (Mendelsohn et al., 2020). To facilitate analysis, we collect a multilingual corpus of $1,340$ Wikipedia biography pages for people in the LGBT community using Wikidata properties and lists of LGBT people from Wikipedia. We additionally use pivot-slope TF-IDF matching to construct a comparison corpus of non-LGBT people. For each page in the target and comparison corpora, we collect article text in English, Russian, and Spanish.

We examine power, agency, and sentiment connotations in the corpus using a verb-centric model

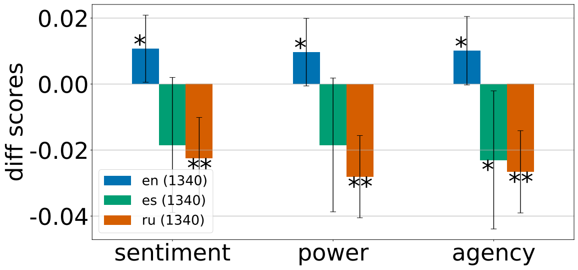


Figure 2.12: Average differences in affective scores in narratives about LGBT people vs. matched control people across languages. In Russian and Spanish, LGBT people are consistently portrayed more negatively, with lower power, and with lower agency, whereas in English, they are portrayed more positively. For all figures, asterisks indicate scores are statistically different from zero (paired t-test, $*$:p<0.05 and $**$:p<0.01) and brackets denote 0.95 confidence intervals. Numbers in the legend or in the title indicate the number of biographies in each group.

similar to the one introduced in §2.1.2, except the model is trained on a new contextualized multilingual data set. Annotations were collected in Russian, English, and Spanish, over sentences and phrases, and a BERT-based classifier is trained over full sentences and phrases (similar to Sentence-level training in Table 2.3). Park et al. (2021) provides details on the data, methodology, and model performance. We report *diff score* as the difference between sentiment/power/agency scores in each each article about an LGBT person and its matched comparison, averaged across all pairs (e.g. "average treatment effect"); a positive score means the articles about LGBT people had a higher affect connotation in aggregate.

Figure 2.12 shows diff scores across the entire corpus. In English, all connotations are significantly positive, whereas in Russian all connotations are significantly negative. All connotations are also negative in Spanish, though only agency is significant. The differences in connotations across languages is surprising, considering that we examine articles about the exact same set of people in all languages.

These trends reflect global perceptions about LGBT people identified by other studies. Data from 2006 and 2011 suggest that many English Wikipedia editors are from the United States, many Russian Wikipedia editors are from Russia, and many Spanish Wikipedia editors are from Spain and (to a lesser extent) other countries, including Argentina, Chile, Netherlands, Mexico, and Venezuela.[24][25] While Wikipedia editor demographics may have since changed, this data shows historical contributions that have been made to Wikipedia and also reflects the countries where these languages are commonly spoken.

While homosexuality was de-criminalized in Russia in 1993, homophobia is still prevalent in the country and can manifest, for instance, in laws against "gay propaganda" (Wilkinson, 2014; Buyantueva, 2018). A report by the Wilkins Institute analyzed survey data from 174 countries to measure their social acceptance of LGBT people (Flores, 2019). Based on data from 2014-2017, the United States ranked 21st (acceptance score of 7.2), while Russia ranked 120th (score of 3.4). Spain ranked highly (rank: 5, score: 8.1), while other Spanish-speaking countries ranked lower: Argentina (23; 6.9), Chile (27; 6.7), Mexico (32; 6.3), Venezuela (39; 5.7), Peru (53; 5.3). The results in Figure 2.12 suggest that these perceptions are reflected in our corpus: LGBT people are portrayed

[24] https://meta.wikimedia.org/wiki/Edits_by_project_and_country_of_origin

[25] https://commons.wikimedia.org/w/index.php?title=File:Editor_Survey_Report_-_April_2011.pdf&page=2

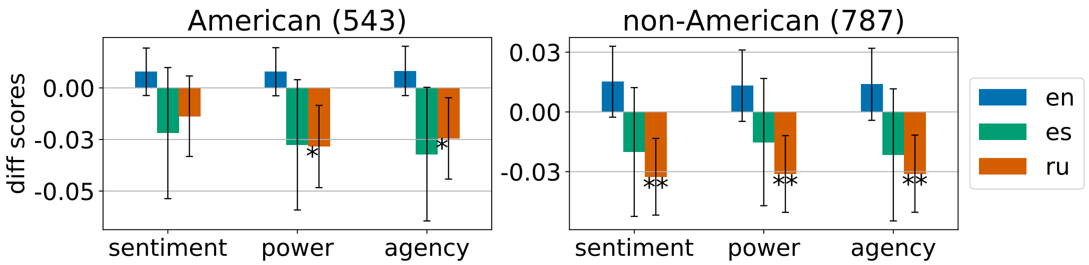


Figure 2.13: Average differences in affective scores for narratives about LGBT and non-LGBT nationality subgroups.

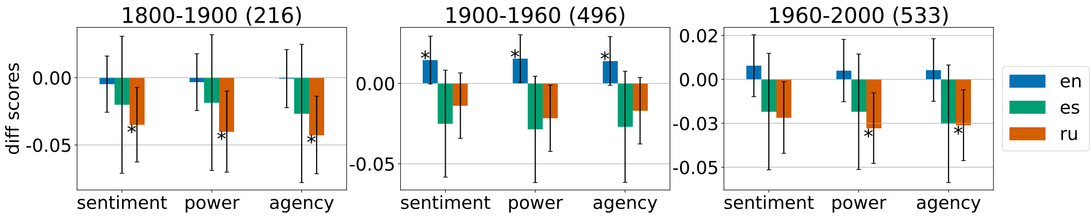


Figure 2.14: Average differences in affective scores for narratives about LGBT and non-LGBT age subgroups.

with more negative connotations in Russian, but not in English. Results are more mixed in Spanish, which is commonly spoken (and edited by on Wikipedia) people from a diverse range of countries.

In order to further examine possible cultural differences and in recognition that sexual orientation does not reflect an individual's entire identity, we divide our corpus according to nationality, birth year, and occupation and test if additional social theories manifest in our data.

Figure 2.13 displays diff scores for each language over American people and non-American people in the corpus. While further subdivisions in nationality could be more informative, we do not report them, out of concern that results over small data sizes could be misleading (e.g. 34 people in our corpus have Russian nationality and 48 have South American nationalities.). In this figure, we investigate the "local heros" hypothesis, which suggests biography pages tend to be longer and more positive for people whose nationality match the language the page is written in (Callahan and Herring, 2011; Eom et al., 2015). Our data does not show evidence of a local bias, in that trends are nearly identical across articles about American and non-American people, even in English, where we might expect to see different portrayals of Americans and non-Americans.

In Figure 2.14, we divide our corpus according to the year of birth of each article's subject. Our primary research question in this figure is if the general increase in global acceptance of LGBT people over time is reflected in our data set (Flores, 2019). Trends in Russian and Spanish do not significantly change across biography pages for people with different birth years. However, in English, portrayals are neutral/insignificantly negative for people born before 1900, but significantly positive for people born 1900-1960 and (insignificantly) positive for people born after 1960. Thus, while not all results are significant, our data does offer some evidence that changing global perceptions of LGBT people are reflected in their English Wikipedia biography pages.

Finally, in Figure 2.15, we subdivide the LGBT corpus by occupation. As in Figure 2.13, we focus only on the two most common occupations identified in our corpus (Entertainer and Artist) in order to ensure sufficient sample size. Survey and behavioral studies have suggested that LGBT peo-

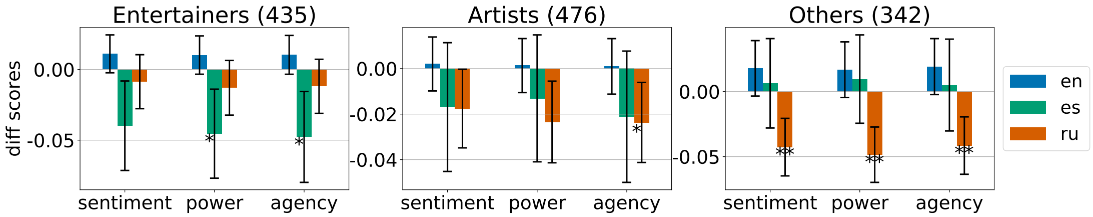


Figure 2.15: Average differences in affective scores for narratives about LGBT and non-LGBT occupation subgroups.

ple are perceived as better suited to some occupations than others (Tilcsik et al., 2015; Clarke and Arnold, 2018). In Figure 2.15, we see little difference in how LGBT people of different occupations are portrayed on Wikipedia in English. However, in Spanish, articles about entertainers have significantly negative connotations, but articles about people of other occupations do not. Conversely, in Russian, while entertainers have near-neutral connotations, people of other occupations, such as politicians, scientists, and activists, are portrayed with significantly negative connotations. While more investigation is needed, our data offers evidence suggesting that perceptions about occupational stereotypes of LGBT people may differ across cultures and languages.

In general, the trends in Figure 2.12 remain consistent across different divisions of the data, with only slight variations. Articles about LGBT people are more negative than comparisons in Russian, but not in English, and results in Spanish are mixed.

2.4 Conclusions and Future Work

In this chapter, we introduce *context affective analysis*, a framework for examining how people are portrayed in narrative next. We show how this framework can be used to examine stereotypes about people with different characteristics in specific settings (e.g. media coverage of the #MeToo movement) as well as in broader corpora (e.g. on Wikipedia or news more generally). We additionally collect several new data sets, including news articles about the #MeToo movement and Wikipedia biography articles with inferred race, gender, and sexual orientation information across multiple languages. The generalizability of this framework allows for examination of a broad range of research questions in diverse types of narrative text.

While this work is motivated in part by the tendency of NLP models to absorb and amplify biases in training data, this chapter focuses on human-written text, and more research is needed to understand what perceptions about people models absorb. For example, does a language model trained on Wikipedia generate text with the same, more, or less stereotypical portrayals of people as the original training data? The methodology developed in this chapter could be used to investigate this type of research question in machine-generated text.

Furthermore, the dimensions of power, agency, and sentiment provide a framework that is relevant to many different social settings, but other narrower dimensions may be more relevant in specific contexts. Researchers may be interested in examining specific stereotypes, like “angry Black woman” (Collins, 1990; Walley-Jean, 2009). Deriving annotated data for every stereotype of interest is time-consuming, expensive, and likely to introduce annotator bias, suggesting additional unsupervised or distantly-supervised methods for identifying specific stereotypes could be useful. Some additional insight could be gained by examining intersections of theses dimensions – are power and sentiment connotations correlated positively for men, but negatively for women? Finally, there are limitations to both the verb-centric (§2.1) and entity-centric approaches (§2.2), in that verb-centric primarily captures verb connotations, whereas entity-centric is less reliable in domain-specific settings. Future research could aim to improve these limitations and develop new evaluation metrics.

Chapter 3

Uncovering Global Opinion Manipulation Strategies

This chapter discusses work previously published in Field et al. (2018) and Tyagi et al. (2020)

While the increasing availability of online text has created unprecedented access to information, it has also raised difficulties in ensuring information quality and opened avenues for disinformation and opinion manipulation campaigns. Trends of computational propaganda have been identified over 80 countries and the manipulation of public opinion over online media is considered "a critical threat to democracy" (Bradshaw et al., 2021). Well-known examples of manipulative or deceptive content include Russian-government-affiliated accounts disingenuously tweeting about U.S. elections and rampant online misinformation about COVID-19 (Starbird et al., 2019).

Public concerns about "fake news" and misinformation have inspired a growing area of NLP research on fact-checking and fake news detection (Hassan et al., 2017; Thorne et al., 2018; Wang et al., 2018; Jiang et al., 2020). These approaches often rely on comparing a *claim* with a trusted information source in order to determine if the claim is supported or refuted. However, actors seeking to manipulate information landscapes or public opinion also use subtle tactics that are much harder to detect, like flooding communication channels with irrelevant information or highlighting particular viewpoints of an event to distract public attention (King et al., 2017; Munger et al., 2018; Rozenas and Stukal, 2019). These strategies do not necessarily involve perpetuating information that is factually incorrect, making them impossible to identify through fact-checking. Identifying and analyzing these types of subtle information operations has largely relied on leaks that reveal ground-truth data (King et al., 2017; Arif et al., 2018).

In this chapter, we develop methods for examining public opinion manipulation strategies by government-associated actors in two settings: Russian newspaper articles spanning 13 years and tweets about the terrorist attack in Pulwama district, Jammu and Kashmir, India, on February 14, 2019. Our approach involves developing computational approaches to identify media manipulation strategies drawn from political science. In the first setting, we focus on: *agenda-setting*—selecting *what* topics to cover—and *framing*—*how* aspects of those topics are highlighted to promote particular interpretations (Entman, 2007; Ghanem and McCombs, 2001). In the second setting, we focus on *polarization*, measuring to what extent actors on Twitter argued for or against de-escalation of conflict.

Importantly, in both cases we develop language-agnostic methodology. While framing, agenda setting, and polarization have been studied previously, research has focused on English-speaking democratic countries, particularly U.S. politics, and most annotated datasets only exist in English (McCombs, 2002; Boydstun et al., 2013; Card et al., 2016; Arif et al., 2018; Badawy et al., 2018; Demszky et al., 2019; Stewart et al., 2017; Mendelsohn et al., 2021). §3.1 describes our methodology and findings for uncovering agenda setting in Russian newspaper articles, §3.2 focuses on framing in the same setting, and §3.3 focuses on polarization in Indian and Pakistani social media. Methodology to expose subtle opinion manipulation campaigns can inform technology platforms in developing mitigation techniques, contribute to research in political science and public policy, and provide insight into how text generation systems can be misused (Brown et al., 2020).

3.1 Agenda Setting in Russian News

3.1.1 Background

> “The media may not be successful much of the time in telling people what to think, but is stunningly successful in telling its readers what to think about” (Cohen, 1963)

Agenda-setting refers to media outlets’ ability to influence the importance of topics through the selection of *what* topics they report on. All media outlets inevitably use a form of agenda-setting: deciding what is “newsworthy” by covering some topics at the exclusion of others, and these decisions can powerfully sway the focus of public opinion (McCombs, 2002; Cohen, 1963). We hypothesize that in countries with weak democratic institutions and in particular, with state-controlled media, the government may actively use agenda-setting to shape public opinion.

We investigate this hypothesis in a corpus of Russian newspaper articles because of intense interest in the way Russia is shaping the global information environment (Van Herpen, 2015). Many Russian media outlets are state-owned or heavily influenced by the government. Our results are based on a corpus of over 100,000 articles from the newspaper *Izvestia* published in 2003–2016. Despite a brief period of autonomy, *Izvestia* has become strongly influenced by the government (Jones, 2002).

Prior work has identified a relationship between negative economic performance in Russia, such as stock market declines, and “selection attribution” in state-controlled media outlets, where negative events are blamed on foreign officials while positive events are credited to domestic officials (Rozenas and Stukal, 2019). We build on these findings and investigate the relationship between economic performance, including that of the Russia Trading System Index (RTSI) and gross domestic product (GDP), and news coverage of foreign events. We primarily investigate coverage of the United States because Russia has seen the U.S. as its main rival since the Cold War, and we expect news coverage of foreign events to focus dis-proportionally on the U.S.

We first establish a strong negative correlation between Russia’s economic situation and the proportion of news focused on the U.S. (§3.1.2). We then show that the correlation is directed: economic indicators precede (and thereby Granger-cause) the increase in foreign news coverage (§3.1.3). The primary contributions of this section include combining economic metrics and text features to automate the identification of agenda-setting as well as exploration of agenda-setting as a media manipulation strategy in a Russian newspaper.

3.1.2 Methodology and Results: Correlations

We compared the salience of news focused on the U.S. with indicators that reflect the economic state of Russia to test our hypothesis: that news coverage of the U.S. is used to distract the public from negative economic events. We first performed an initial, simplistic study of this agenda-setting strategy. We define *U.S. coverage* as the ratio of *Izvestia* articles that mention the U.S. at least twice to the total number of articles in any given time slice (in our initial study, a year). We show in Figure 3.1 the U.S. coverage plotted against Russian GDP, in an annual resolution. We find a strong negative Pearson's correlation (r=-0.83): mentions of the U.S. in *Izvestia* increase as economic indicators deteriorate. The one exception to this trend is 2008, during which there was a high amount of U.S. news coverage and the Russian GDP peaked. This year coincides with both the U.S. financial crisis and the Obama-McCain Presidential election, which would explain a focus on U.S. events regardless of the Russian economic situation.

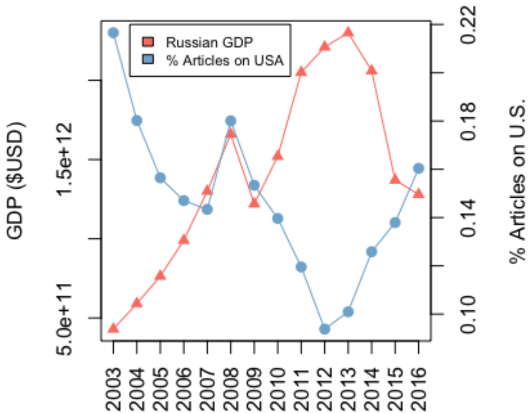


Figure 3.1: Proportion of articles that mention the U.S. at least twice (blue) and Russian GDP (red), 2003–2016.

We next extend these preliminary results in several ways. First, we refine the definition of *U.S. coverage* by using two metrics: **article level,** the number of articles that mention the U.S. at least twice normalized by the total number of articles in the time slice; and **word level,** the frequency of the occurrences of the U.S., normalized by the total number of words in the time slice. Second, we compare these metrics to *two* economic indicators: GDP (in USD) and the index of the Russian stock market (RTSI), in rubles.[1] Third, we refine the time-resolution and use yearly, quarterly, and monthly time slices.

Table 3.1 reports the correlations between the two metrics of U.S. coverage and (monthly, quarterly, and yearly) RTSI and GDP values. At all levels, there are strong negative correlations between the proportion of news focused on the U.S. and economic state.

Level	Article	Word
RTSI (Monthly, rubles)	-0.54	-0.52
GDP (Quarterly, USD)	-0.69	-0.65
GDP (Yearly, USD)	-0.83	-0.79

Table 3.1: Pearson's correlation between news coverage of the U.S. and economic indicators.

[1] Stock market values were obtained from the Moscow exchange website. GDP values were obtained from OECD.

These negative correlations between *U.S. coverage*, measured by counting mentions of the U.S. in *Izvestia*, and the state of the Russian economy indicate the possibility of intentional agenda-setting by the Russian government.

3.1.3 Methodology and Results: Granger Causality

To conclude that agenda-setting may be ocurring, it is necessary to show that these correlations are in fact directed: a change in the economy is followed by a change in U.S. news coverage. To investigate directionality, we employ Granger causality (Granger, 1988). The key concept behind Granger causality is that cause precedes effect. Thus, a time series X is said to Granger-cause a times series Y if past values x_{t-i} are significant indicators in predicting y_t. First, we computed the article-level (a_t) and word-level (w_t) metrics at a monthly granularity from 2003 to 2016; we also extracted the RTSI monthly close price (in USD) for the same time period (r_t). We then calculated the percentage change of these series as: $C(w_t) = \frac{w_t}{w_{t-1}} - 1$, and equivalently calculated $C(a_t)$ and $C(r_t)$. By taking the percent change of both series, we control for long term trends (e.g., stock markets tend to trend upwards over time), and instead focus on short-term relations: does a change in the economy directly precede a change in news coverage?

We computed Granger causality between $C(w_t)$ and $C(r_t)$ by fitting a linear regression model:

$$C(w_t) = \sum_{i=1}^{m} \alpha_i (C(w_{t-i})) + \sum_{j=1}^{n} \beta_j (C(r_{t-j}))$$

where m and n denote how far back in time we look (denoted as m-lag or n-lag). We can say that r_t Granger-causes w_t if we find that β is significantly different from zero.

Tables 3.2 and 3.3 report the results. A p-value ≤ 0.05 indicates significance; thus we find 1-lag RTSI values Granger-cause coverage of U.S. news by both the word-level and article-level metrics. Importantly, the r_{t-1} coefficient is negative, which indicates that a decline in the stock market is followed by an increase in U.S. news coverage. In the 2-lag analysis, the r_{t-2} values are not significant, which suggests that the changes in news coverage follow changes in the stock market within one month.[2] For completeness, we also computed Granger causality in the reverse direction: i.e., does a change in U.S. news coverage Granger-cause a change in the stock market? As expected, we found no significant results.

	1-Lag		2-Lag	
	α; β	p-value	α; β	p-value
w_{t-1}	-0.233	0.003	-0.320	0.00005
w_{t-2}	-	-	-0.301	0.0001
r_{t-1}	-0.352	0.0334	-0.369	0.024
r_{t-2}	-	-	-0.122	0.458

Table 3.2: Granger causality between % change in RTSI and frequency of USA (word level).

Overall, these results provide evidence of agenda-setting in *Izvestia*. In examining 13 years of articles, we find significant negative correlations between the state of the Russian economy and the

[2] We computed Granger causality at a quarterly and yearly level and found no significant causal relationship. This result is unsurprising; the monthly analysis suggests trends in news coverage are largely driven by the previous month, so we would not expect causality at a quarterly or yearly level.

	1-Lag		2-Lag	
	α; β	p-value	α; β	p-value
w_{t-1}	-0.222	0.005	-0.290	0.000289
w_{t-2}	-	-	-0.270	0.000634
r_{t-1}	-0.311	0.035	-0.329	0.0267
r_{t-2}	-	-	-0.091	0.543

Table 3.3: Granger causality between % change in RTSI and frequency of USA (article level).

amount of U.S. news coverage. We additionally find that economic decline precedes the increase in U.S. news coverage, when examined at a monthly level.

3.2 Framing in Russian News

3.2.1 Background

While §3.1 provides evidence of *agenda-setting* in Russian news articles, this analysis does not fully reveal why discussing the U.S. during economic downturns is an effective manipulation strategy–what are these articles saying about the U.S.? We hypothesize that *framing* can further our understanding of why Russian media focuses on the U.S. during economic downturns. While agenda-setting broadly refers to what topics a text covers, framing refers to which *attributes* of those topics are highlighted, often to promote particular interpretations (Entman, 2007; Ghanem and McCombs, 2001).

Several aspects of framing make the concept difficult to analyze. First, just defining framing has been “notoriously slippery” (Boydstun et al., 2013). Frames can occur as stock phrases, i.e. “death tax” vs. “estate tax”, but they can also occur as broader associations or sub-topics (Tsur et al., 2015; McCombs, 2002). Frames also need to be distinguished from similar concepts, like sentiment and stance. For example, the same frame can be used to take different stances on an issue: one politician might argue that immigrants boost the economy by starting new companies that create jobs, while another might argue that immigrants hurt the economy by taking jobs away from U.S. citizens (Baumer et al., 2015; Gamson and Modigliani, 1989). Finally, unlike classification tasks where each article is assigned to a single category, most articles employ a variety of frames (Ghanem and McCombs, 2001).

Recent work has attempted to address these conceptual challenges by defining broad framing categories. The Policy Frames Codebook defines a set of 15 frames (one of which is “Other”) commonly used in media for a broad range of issues (Boydstun et al., 2013). In a follow-up work, the authors use these frames to build The Media Frames Corpus (MFC), which consists of articles related to 3 issues: immigration, tobacco, and same-sex marriage (Card et al., 2015). About 11,900 articles are hand-annotated with frames: annotators highlight spans of text related to each frame in the codebook and assign a single “primary frame” to each document. However, the MFC, like other prior framing analyses, relies heavily on labor-intensive manual annotations.

The primary automated methods have relied on probabilistic topic models (Tsur et al., 2015; Boydstun et al., 2013; Nguyen et al., 2013; Roberts et al., 2013). Although topic models can show researchers what themes are salient in a corpus, they have two main drawbacks: they tend to be corpus-specific and hard to interpret. Topics discovered in one corpus are likely not relevant to a different corpus, and it is difficult to compare the outputs of topic models run on different corpora.

Other automated framing analyses have used the annotations of the Media Frame Corpus to predict the primary frame of articles (Card et al., 2016; Ji and Smith, 2017), or used classifiers to identify language specifically related to framing (Baumer et al., 2015). Importantly, all of these methods focus exclusively on English data sets. While unsupervised methods like topic models can be applied to other languages, any supervised method requires annotated data, which does not exist in other languages.

In this section, we present a new method for analyzing frames and evaluate it quantitatively through hand-annotations and qualitatively through a series of examples. We then use this method to contextualize strategies of media manipulation in the *Izvestia* corpus.

3.2.2 Methodology

Our goal is to develop a method that is easy to interpret and applicable across-languages. In order to ensure our analysis is interpretable, we ground our method using the annotations of the Media Frames Corpus. However, because the MFC is entirely in English and our test corpora is in Russian, we cannot use a fully supervised method. Instead, we use the MFC annotations to derive lexicons for each frame, which we then translate into Russian. We use query-expansion to reduce the noisiness of machine translation and make the lexicons specific to the *Izvestia* corpus, rather than specific to the MFC. We evaluate the derived lexicons in English and in Russian. Finally, we use these lexicons to analyze frames in *Izvestia* and identify strategies of media manipulation. Our method allows for in-depth analysis by identifying primary and secondary frames in a document and specific words that signify frames.

Generating framing lexicons Although our primary test corpus is in Russian, we also use English test corpora for evaluation; thus, we describe our method as applicable to either language. First, we use the MFC annotations to derive a lexicon of English words related to each frame in the Policy Frames Codebook. For a given frame F we measure pointwise mutual information (Church and Hanks, 1990) for each word in the corpus as:

$$I(F,w) = log\frac{P(F,w)}{P(F)P(w)} = log\frac{P(w \mid F)}{P(w)}$$

We estimate $P(w|F)$ by taking all text segments annotated with frame F, and computing $\frac{Count(w)}{Count(allwords)}$. We similarly compute $P(w)$ from the entire corpus. We then use the 250 words with the highest $I(w,F)$ as the base framing lexicon for frame F, denoted F_{base}. We discard all words that occur in fewer than 0.5% of documents or in more than 98% of documents.

Translation and extension of framing lexicons Next, we use query-expansion to alter F_{base}, with the goal of generalizing the lexicon. Without this step, our lexicons are biased towards words common in English news articles, particularly words specific to the 3 policy issues in the MFC.

When our test corpus is in a different language (i.e. Russian), we use Google Translate to translate F_{base} into the new language. We restrict our vocabulary to the 50,000 most frequent words in the test corpus.

Then, to perform the query-expansion, we train 200-dimensional word embeddings on a large background corpus in the test language, using CBOW with a 5-word context window (Mikolov et al., 2013). We compute the center of each lexicon, c, by summing the embeddings for all words in the lexicon. We then identify up to the K nearest neighbors to this center, determined by the cosine distance from c, as long as the cosine distance is not greater than a manually-chosen threshold (t).[3] We again filtered the final set by removing all words that occur in fewer than 0.5% of documents or in more than 98% of documents.

The final lexicons contain between 100 and 300 words per frame. Table 3.4 depicts a few examples of lexicon words extracted from the MFC (F_{base}), and words in our final lexicons adapted to the *Izvestia* corpus (F_{rus}). We can observe that words in F_{rus} are closely related to words in F_{base}, but also specific to Russian culture and politics.

We consider a document to employ a frame F if the document contains at least 3 instances of a word from F's lexicon. We assign the primary frame of a document to be its most common frame, determined by the number of words from each framing lexicon in the document.[4]

F_{base}	F_{rus}
Political	
republican-controlled	bills
filibuster	conservative
gubernatorial	parlimentary
Economic	
cents	deductions
holdings	tax
profitable	fines
Public Sentiment	
gallup	activism
demonstrators	facsim
rallied	vote

Table 3.4: Example lexicons extracted from the MFC and transfered to the *Izvestia* corpus.

3.2.3 Evaluation

We evaluate the English lexicons using annotated data from the MFC. For the Russian lexicons, since we do not have annotated Russian data, we instead conduct an annotation task. These evaluation metrics determine how well our method captures which frames are present in a text. Finally, we also qualitatively compare our method to existing methods for framing analysis, specifically topic models.

English Evaluations We first evaluate our lexicons on two tasks using the MFC annotations: primary frame identification and identification of all frames in a document.

Primary frame identification is a 15-class classification problem. Two prior studies evaluate models on this task: Card et al. (2016) and Ji and Smith (2017). Following these studies, we evaluate our model using 10-fold cross-validation on only the "Immigration" subset of the MFC. We

[3] When the test corpus is in English, we set t to 0.4 and K to 500 and we add the identified neighbors to F_{base}. When our test corpus is in Russian, we choose to discard our base lexicon, to prevent the final lexicons from being too U.S.-specific. Instead, we set t to 0.3 and K to 1000, which increases the number of neighbors identified, and we keep only these neighbors in the final lexicon.

[4] We do not generate a lexicon for the "Other" frame, and instead assign a document's primary frame as "Other" only if it does not contain at least 3 words from any framing lexicon. Throughout this process, we use small subsets of the "tobacco" articles for parameter tuning.

use 9 folds to generate framing lexicons and the 10th fold to evaluate. To train word embeddings, we use the entire MFC corpus combined with over 1 million New York Times articles from 1986 - 2016 (Fast and Horvitz, 2017). Table 3.5 shows the accuracy of our model. Our results outperform Card et al. (2016) and are comparable to Ji and Smith (2017). Furthermore, unlike prior methods, our method is able to transfer to different domains and languages without needing further annotated data.

Ji and Smith (2017)	58.4
Card et al. (2016)	56.8
Our model	57.3

Table 3.5: Accuracy of primary frame classification.

However, our main interest is in measuring the salience of frames in general, not merely focusing on the primary frame. Thus, we also use our lexicons to identify the presence of any frames in a document. As the MFC has multiple annotators, we define a frame to be present in a document if any annotator identified the frame, and use this as gold standard test data. In evaluating our lexicons, we consider a frame to be present in a document if the document contains at least 3 tokens from the frame's lexicon.

To the best of our knowledge, identifying all frames in a document is a new task that was not attempted in prior work. Thus, we use a logistic regression model with bag-of-word features as a standard baseline. As above, we evaluate using 10-fold cross validation on the "Immigration" subset of the MFC. Table 3.6 shows that our method outperforms the baseline, with the exception of 2 frames, even though the baseline is fully supervised, whereas our method is distantly supervised. We note that the poorest performing frames, "External Regulation and Reputation" and "Morality" are the frames which are least common in this subset of the data – each frame occurs in fewer than 500 articles. When we run the same 10-fold cross validation evaluation on the "Samesex" subsection of the MFC, where the "Morality" frame occurs in over 1000 articles, we achieve a higher F1 score (0.65).

	Ours	Baseline
Capacity & Resources	**0.53**	0.48
Crime & Punishment	**0.78**	0.76
Cultural Identity	0.57	**0.62**
Economic	**0.69**	0.67
External Regulation	0.25	**0.47**
Fairness & Equality	**0.50**	0.44
Health & Safety	**0.58**	0.53
Legality & Constitutionality	**0.80**	0.76
Morality	**0.31**	0.25
Policy Prescription	**0.72**	0.69
Political	**0.80**	0.77
Public Sentiment	**0.54**	0.47
Quality of Life	**0.65**	0.63
Security & Defense	0.63	0.63

Table 3.6: F1 scores for identification of all frames in a document.

Russian Evaluations Next, we evaluate the quality of our method on the Russian data set. Unlike in English, we do not have frame-annotated data in Russian. We instead performed the intruder detection task, an established method for evaluating topic models (Chang et al., 2009). For each frame F we randomly sampled 5 words from the framing lexicon F_{rus} and 1 word from the lexicon of a different frame, which has no overlap with F_{rus}. We then presented two (native Russian speaking) annotators with the frame heading and the set of 6 words, and asked them to choose which word did not belong in the set. We evaluated 15 sets or 75 words per frame.

Framing can be subjective, and we do not necessarily expect annotators to interpret frames in the same way. We calculate two forms of accuracy: "soft", whether *any* annotator correctly identified the intruder; and "hard", whether both did. We also report average precision as defined in Chang et al. (2009), i.e. the average number of annotators that correctly identified the intruder, averaged across all sets.

We briefly summarize results here and report them fully in Field et al. (2018). All accuracy scores are significantly better than random guessing, and no soft accuracy falls below 60%. Only two frames have an average accuracy $\leq 60\%$, "Fairness and Equality" and "Morality", both very abstract concepts. In these frames, we also see a larger difference between hard and soft metrics, which reflects the subjectivity of framing. The MFC annotators sometimes disagreed on the correct annotations, even after discussing their disagreements (Boydstun et al., 2013). Thus, we can attribute some of the differences between hard and soft accuracies to this subjectivity.

Qualitative Comparison to Structured Topic Models We also qualitatively compared the information our framing lexicons provide with information provided by a Structured Topic Model (STM) (Roberts et al., 2013). We find that our approach is better able to capture frames the way a reader might conceptualize them, whereas topic models are useful for finding fine-grained corpus-specific topics.

Topic models are a common way to analyze frames in a text (Nguyen et al., 2013); the STM specifically allows correlation between topics and covariates. We trained an STM with 10, 15, 20, 25, and 50 topics on U.S.-focused articles in the *Izvestia* corpus, including publication date (month and year) as a covariate. We selected the 20 topic model as having the most coherent topics. Throughout this section, we refer to topics using their most representative words as determined by the "Lift" metric (Roberts et al., 2013).

We randomly selected a sample document for each primary frame to investigate. The framing lexicons are able to connect corpus-specific vocabulary to higher-level concepts. For example, an article describing movies about the U.S. prison facility at Guantanamo Bay has two main STM topics: [laden, sentence, prison] and [author, viewer, filming]. Similarly, the framing lexicons identify 'Cultural Identity" as the primary frame. However, a secondary frame in the document is "Morality", captured by words: writer, form, Christ, art. While both the STM and the framing lexicons capture major details of the article, the framing lexicons additionally tie the article to morality, because words like "art" in this corpus are often signs of a moral framework.

Nevertheless, when the STM identifies a topic similar to a frame, we find correlations with the related lexicon, i.e. there is a 0.75 correlation between the frequency of words in the Legality, Constitutionality, Jurisdiction lexicon and the monthly average proportion of each document assigned to the topic [yukos, bill, legislation].

Additionally, the framing lexicons tend to have higher precision in identifying relevant articles than the STM. Topics are commonly identified by their most probable words, which may not occur at all in documents associated with the topic. For example, the STM assigns an article about smoking policies in the U.S. to 3 main topics: [laden, sentence, prison], [kosovo, falcons, because], [author, viewer, filming], none of which are closely related to the article. In contrast, because assignments to the framing lexicons are made directly from words in the lexicon, we can be confident that articles assigned to each frame have words from the actual lexicon, and are very likely related to the frame. The framing lexicons assign the primary frame as “Policy” for this article, which is a good fit. Neither method captures that the article is also related to health.

Finally, the STM is useful for finding fine-grained topics, beyond the Policy Frames Codebook. For example, we find a “sports” topic: [match, nhl, team]. These topics tend to be corpus-specific and more concrete than the framing lexicons: no STM topic captures “Quality of Life”.

3.2.4 Identifying Media Manipulation

We first use the generated framing lexicons to determine which frames are frequently associated with the U.S. We then break the frames into finer-categories and manually look at sample articles to determine why associating these frames with the U.S. constitutes a media manipulation strategy. We find that as the stock market declines, not only is news focused more on the U.S., but also emphasizes threats to the U.S.

Salient frames To estimate which frames are associated with the U.S. we compute *normalized pointwise-mutual information* (nPMI) between the U.S. and each frame F[5] by mapping the mutual information score onto a [-1,1] scale. A value of 1 represents complete co-occurrence; a value of 0 represents complete independence. By using nPMI, we measure which frames are *overrepresented* in U.S.-focused news, as compared to other news.

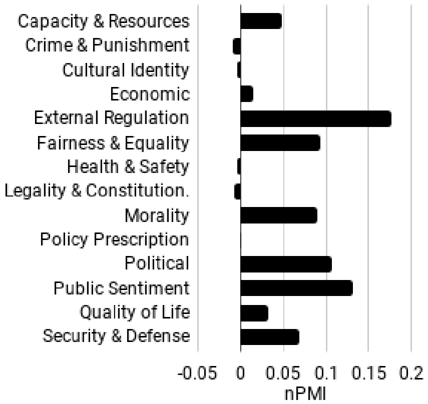


Figure 3.2: nPMI between U.S. and each frame.

Figure 3.2 shows the nPMI score between the U.S. and each frame for all articles in our corpus. As any news article about the U.S. is by definition externally focused, the frame with the strongest association is unsurprisingly “External Regulation and Reputation”. Other frames with strong

[5] As above, we consider an article to be U.S.-focused if it mentions the U.S. at least 2 times, and we consider an article to employ frame F if it uses at least 3 words from F's lexicon.

associations include “Morality”, “Political”, “Public Sentiment”, and “Security and Defense”. These frames demonstrate what type of news events in the U.S are reported in Russia. As an example, we look at an article that uses a combination of these frames. The article describes cooling relations between Russia and the U.S. It explains that anti-Russian sentiment will be prevalent in the U.S. during upcoming elections, when politicians on both sides will play the “Russia card”. It ultimately attributes the cooling relations to a mismatch of values and ideology between the two nations. The framing lexicons well-capture the numerous themes in this article. Specifically, the frames identified and the related framing lexicon words are:

Political: electorate, election, former, pre-election, political scientists, congress, president, post, bush

Public Sentiment: elections, campaign, pre-election, democrats, republicans

External Regulation and Reputation: west, war, former, washington, politics, summit, exacerbation, west, decision, bush, president

Morality: peace, sins, ideals, love, values

Fairness and Equality: politics, love, values

This article uses several strategies to promote unity in Russia and actively separate Russia from Western culture, including criticizing American politics and emphasizing a difference of values. Russian articles use a combination of frames to describe the U.S., which demonstrates the importance of looking at all frames in a document, rather than just the primary frame. In the following sections, we provide additional examples demonstrating how combinations of frames can generate anti-U.S. sentiment.

Salient words within frames We expect different aspects of frames to be foregrounded during economic upturns than in downturns. To investigate these differences, we define a set of months M_t^+, as the 10% of months where RTSI showed the greatest growth, and a corresponding set M_t^- where RTSI showed the greatest decline. We then take M_{t+1}^+ as the month directly following every month in M_t^+, and we similarly define M_{t+1}^-. From the analysis in §3.1.3, we expect media manipulation strategies to decline from M_t^+ to M_{t+1}^+, and increase from M_t^- to M_{t+1}^-. For each frame, we take the subset of U.S.-focused articles that use the frame. Then, we identify words that are overrepresented or underrepresented from M_t^+ to M_{t+1}^+ and from M_t^- to M_{t+1}^- using log odds with a Dirichlet prior (Jurafsky et al., 2014; Monroe et al., 2008), specifically taking the 500 words with the largest increase in salience from M_t^- to M_{t+1}^- intersected with the 500 words with the largest decrease in salience from M_t^+ to M_{t+1}^+. Thus, for each frame, we identify words which become more common after a stock market downturn and become less common after a stock market upturn. We refer to these words as AgendaLex.

We found the Security and Defense AgendaLex and the Crime and Punishment AgendaLex to be surprisingly coherent, both containing words related to terrorism and countries enemy to the U.S., including bombs, missiles, Guantanamo, North Korea, Iraq, etc. We found a correlation of -0.49 between the frequency of words from the Security and Defense AgendaLex in U.S.-focused articles and the RSTI (-0.49). A 1-lag Granger causality test (to what extent does a change in RTSI Granger-cause a change in the prevalence of the Security and Defense AgendaLex?) has a p-value of 0.0051. As the stock market declines, not only does the news focus more on the U.S., the news focuses specifically on terrorists and other enemies to the U.S. In the next section, we refine this conclusion by looking at sample articles.

Examples of framing during downturns By reading sample articles from months just after stock market downturns that used words from the Security and Defense lexicon and AgendaLex, we identified three common strategies for distracting Russian citizens from negative economic events: villainizing the U.S., describing threats to the U.S., and promoting the Russian military over the U.S. military.

First, some articles focus on immoral actions of the U.S. military, describing U.S. troops as "Nazi", or U.S. campaigns in Iraq as "barbaric" or causing "horror and outrage throughout the world". Others discuss Guantanamo Bay, employing the "Morality" or "Legality, Constitutionality, Jurisdiction" frames. By portraying the U.S. government negatively, actions of the Russian government appear positively by comparison. Promoting unity by presenting an external enemy is a well-studied political strategy.

Second, many articles, often in passing, described threats to the U.S. An article about terror attacks in Paris mentions U.S. involvement with words from the Crime and Punishment and Security and Defense lexicons: terrorist, terrorists, special services. An article about the conflict between Israel and Palestine describes increased security in the U.S. An article about a U.S. military operation refers more directly to threats to the U.S. by claiming the killed terrorist will simply be replaced "and everything will start afresh - explosions, chases, roundups...unlucky businessmen, successful terrorists". The articles portray the U.S. as an unsafe place to live, making Russia seem like a preferable home.

A third type of article also presents Russia as safe by downplaying U.S. military threat: "the missile defense system of the USA does not pose a real threat to Russia's strategic nuclear forces." or describing the growth of Russian technology compared to 'impotent' American counterparts.

Conclusions Our method for analyzing frames contributes to the growing body of work on automated-framing analysis (Nguyen et al., 2013; Boydstun et al., 2013; Card et al., 2016; Baumer et al., 2015). While past work uses fully-supervised methods, which are not applicable to languages lacking training data, or unsupervised topic models, which can be difficult to interpret, we take a semi-supervised approach: using statistical metrics and word embeddings to generate corpus-specific lexicons based on common frameworks. When combined with methods for detecting agenda-setting presented in §3.1, we show that natural language technology, in addition to its ability to address overt manipulation strategies like "fake news" and censorship, has the potential to shed light on more subtle political manipulation strategies, specifically distraction. We offer a way to define these strategies by drawing on social science theories of agenda-setting and framing, combining them with a novel methodology for cross-lingual projection of framing annotations. We investigate how the resulting frames are used in the Russian newspaper *Izvestia*, and show that it reports on negative events in the U.S. as a way of distracting from economic downturns in the Russian economy.

3.3 Polarization on Indian and Pakistani Social Media

3.3.1 Background

While §3.1 and §3.2 focus on a government-influenced newspaper outlet in Russia, in this section we examine content posted on Twitter around a specific incident: a terrorist attack in Pulwama, Kashmir, which led to retaliatory airstrikes and rising tensions between Indian and Pakistan from February 14, 2019 to March 4, 2019. Like mainstream media, social media can strongly influence public opinion (Bradshaw and Howard, 2018). Using automated methods to analyze social media offers a way to understand the type of content users are exposed to, the positions taken by various users, and the agendas pursued through coordinated messaging across entire platforms. However, prior computational social science research on polarization has focused primarily on U.S. politics, and much attention has focused on the influence of Russian or Chinese state actors (Arif et al., 2018; Badawy et al., 2018; Demszky et al., 2019; Golovchenko et al., 2018; King et al., 2017; Le et al., 2019; Starbird et al., 2019; Stewart et al., 2017).

In contrast, we focus on polarizing social media content in India and Pakistan following the terrorist attack in the Pulwama district, Jammu and Kashmir, India, on February 14, 2019. We primarily investigate: *to what extent did entities on social media advocate for or against escalating tensions?*. India and Pakistan are both nuclear-armed countries and have a decades-long history involving multiple armed conflicts. The Pulwama attack in 2019 was followed by an escalation of tensions between these two nations that nearly approached full-fledged war (Feyyaz, 2019; Palakodety et al., 2020; Pandya, 2019). Moreover, the relationship between these countries is an important agenda for political parties in both India and Pakistan. India has two primary political parties: the Indian National Congress (INC), which was dominant in the early 21st century, and the Bharatiya Janata Party (BJP), which rose to prominence on a populist and nationalist platform in 2014 and has been in power since (McDonnell and Cabrera, 2019). BJP is well-known for promoting nationalism, and journalists and community members have speculated that BJP party members advocated for escalating conflict during this time, because conflict with Pakistan could increase Prime Minister Modi's chances of winning the upcoming elections in April (Mishra, 2019; McDonnell and Cabrera, 2019).[6] Given this context, we first examine the tweets and communication patterns of general users in order to understand how polarizing the attack was and to what extent users with different viewpoints may have interacted with each other. We then examine the social media messaging of political party members and how it changed over the sequence of events in order to uncover possible political agendas.

Our core methodology uses a network-based label propagation algorithm to quantify the polarity of hashtags along specified dimensions: Pro-India vs. Pro-Pakistan and Pro-Aggression vs. Pro-Peace. We then aggregate the hashtag-level scores into tweet-level and user-level scores, e.g. the polarity of a given user on a given day. Unlike methodology that assumes users' opinions do not change (Darius and Stephany, 2019; Darwish, 2019; Weber et al., 2013), focuses on binary stances (Borge-Holthoefer et al., 2015), or requires in-language annotations and feature-crafting (Magdy

[6] https://www.forbes.com/sites/kenrapoza/2019/02/27/indias-fight-with-pakistan-seen-lifting-modis-election-chances/#1df2795a397c
https://www.theweek.in/news/india/2019/02/28/imran-s-party-slams-modi-on-escalating-indo-pak-tension.html
https://www.washingtonpost.com/opinions/2019/03/04/after-pulwama-indian-media-proves-it-is-bjps-propaganda-machine/

et al., 2016), our methodology allows us to analyze degrees of polarization in a multilingual corpus and how they change over time.

3.3.2 Methodology

We collected tweets that contained hashtags related to these events, where we include both hashtags more likely to be used by Pro-India users (e.g., IndiaWantsRevenge) and hashtags more likely to be used by Pro-Pakistan users (e.g., PakistanZindabad). Our collected data set contains 2.5M unique tweets (including retweets) from 567K users that use 67K unique hashtags. All tweets occurred between February 14^{th} and March 4^{th}. The data contains a mix of languages including English, Urdu, and Hindi, and many users use multiple languages in the same tweet. While some tweets express neutral opinions, others contain incendiary language.

We develop a method to assign a polarity score to an aggregate group of tweets, and we analyze how polarities change over time for different groups of users. For instance, given pole A (e.g., Pro-Pakistan) and pole B (e.g., Pro-India), we aggregate all tweets by a given user and assign the user a polarity score between [a, b], where a score close to a indicates the user more likely supports A and a score close to b indicates the user more likely supports B. We could also aggregate only tweets by the user on one day and determine the user's Pro-A/Pro-B polarity on that day.

In the absence of annotated data, we use a weakly supervised approach. First, for pole A, we hand-select a small seed set of hashtags that are strongly associated with A, and we equivalently hand-select a hashtag set S_B. Then, we use S_A and S_B to infer polarity scores over a larger lexicon hashtags $\mathcal{V}$, where each $w \in \mathcal{V}$ is assigned a score in $[a, b]$. Finally, we estimate the polarity of an aggregated set of tweets by averaging the inferred polarity scores for all $w \in \mathcal{V}$ used in those tweets. In order to propagate the hand-annotated labels in S_A and S_B to the larger lexicon $\mathcal{V}$, we develop *network-based hashtag propagation*, which makes the assumption that hashtags with similar polarities are commonly used in the same tweets. Tyagi et al. (2020) provides additional details on the hashtag propagation algorithm, as well as details on data collection and evaluation.

3.3.3 Analysis of Polarization

We investigate multiple aspects of our data set, including network structure, polarities of various entities, and changes over time. Based on prior work suggesting that political entities in India and Pakistan may use social media to influence public opinion (Ahmed et al., 2016; Antil and Verma, 2019; Kumar et al., 2016; Singh et al., 2019), we pay particular attention to the Twitter accounts of politicians as a method for uncovering political agendas.

Table 3.7: Overall polarities of users and tweets.

Position	Unique Users	Total Tweets	Position	Unique Users	Total Tweets
Pro-India	125K (23%)	1.16M (46%)	Pro-Aggression	78K (14%)	626K (25%)
Pro-Pakistan	117K (20%)	764K (30%)	Pro-Peace	252K (45%)	1.48M (59%)
Unclassified	325K (57%)	578K (23%)	Unclassified	237K (40%)	351K (16%)

What are the overall polarities of our data set? In Table 3.7, we obtain polarity scores for each user and tweet and then ternarize them into Pro-India/Pro-Pakistan/Unclassified and Pro-

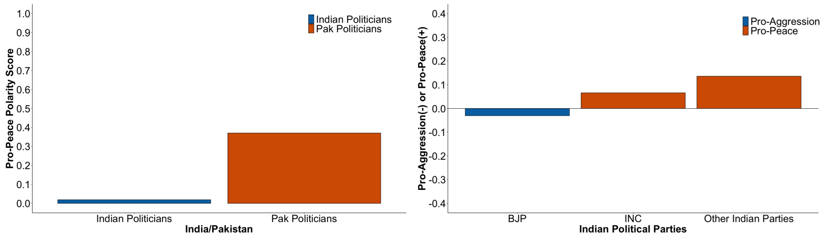


Figure 3.3: Aggregate Pro-Peace and Pro-Aggression polarities of the most popular Indian (33/78) and Pakistani (36/66) politicians in our data set (left) and of members of Indian political parties (right).

Peace/Pro-Aggression/Unclassified. At the user level, the classified accounts are approximately balanced between Pro-India and Pro-Pakistan. However, at the tweet level, the classified data contains a high percentage of Pro-India tweets, suggesting Pro-India users tweeted about this issue more prolifically. Further, there is a much higher percentage of Pro-Peace users than Pro-Aggression users. This pattern also holds at the tweet level, where only a small percentage of tweets are unclassified.

How polarized were different political entities? We investigate the polarities projected by different political entities: specifically BJP politicians (currently in power in India), INC politicians (largest opposition party), other Indian politicians, and Pakistani politicians. We used the Socialbakers.com platform to obtain the Twitter handles of the 100 most followed politicians in India and Pakistan. Our data contained tweets from 66 Pakistani and 78 Indian politicians, and our hashtag model inferred scores for 36 Pakistani and 33 Indian politicians. Figure 3.3 (left) reports aggregate polarity scores over all tweets from these politicians. Pakistani politicians were predominantly Pro-Peace, while Indian politicians expressed mixed polarities, yielding a near neutral score.

We then examined a broader set of Indian politicians, subdivided by political party based on a list of members running for parliament elections in 2019 (Kumaraguru et al., 2019). Out of the 1,360 Twitter handles in the list, our data set contained activity from 316 BJP accounts, 281 INC accounts, 204 other Indian party accounts.

Figure 3.3 (right) shows the overall polarities, aggregated from all tweets by verified members of each party. Strikingly, members of the BJP party are positioned as much more Pro-Aggression than the members of either the INC or other parties, and the party overall obtains a Pro-Aggression polarity score. This score is not dominated by 1-2 strongly polarized members of the party: if we aggregate the polarity scores by individuals instead of by party, 15% of BJP members had net Pro-Aggression scores and 13% had net Pro-Peace scores, in comparison to 10% Pro-Aggression/25% Pro-Peace for INC, and 6% Pro-Aggression/29% Pro-Peace for other parties. These results support observations made by journalists and community members about the polarities project by the BJP party during these events.

How did polarization change over time? Figure 3.4 shows how this polarity changed over the two-week period of events: we infer a Pro-Peace/Pro-Aggression polarity score for all tweets posted by members of the specified political subgroup, and we plot the average score across tweets posted each day.

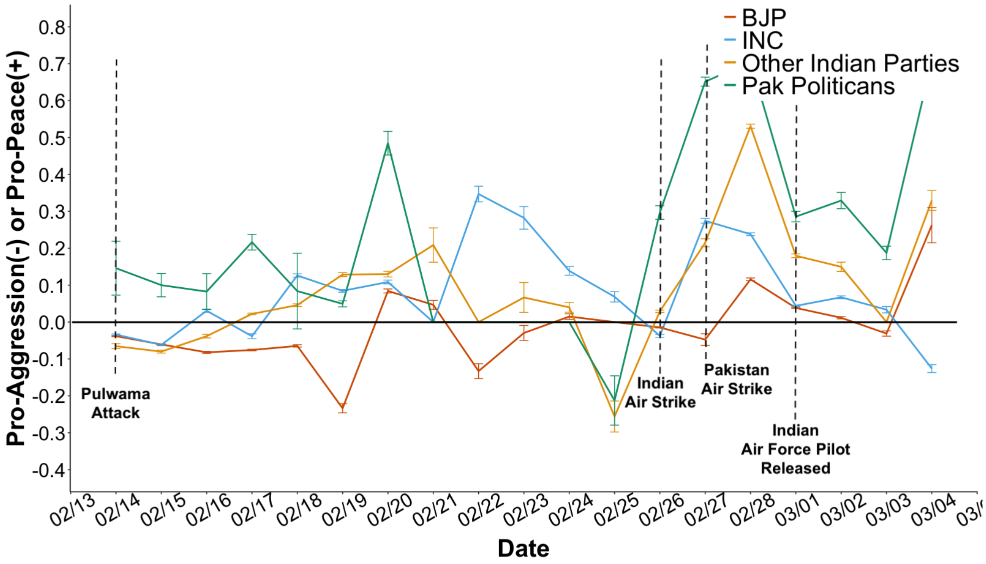


Figure 3.4: Daily Pro-Peace/Pro-Aggression positions of political entities. Negative values denote net Pro-Aggression polarity and positive values denote net Pro-Peace. The error bars represent ±1 standard deviation.

Immediately following the initial attack on 2/14, the tweets from all Indian political party members are inclined towards Pro-Aggression, suggesting initial outrage. However, over the next few days, while tweets from INC and other Indian political party members switch towards Pro-Peace, tweets from BJP politicians remain consistently Pro-Aggression. There is high volatility between 2/20 and 2/26. However, there was a much lower volume of tweets about the Pulwama incidents during this time period, and we do not believe these fluctuations are meaningful. The volume of tweets increases once again following retaliatory Indian (2/26) and Pakistani (2/27) airstrikes. Tweets by Pakistani politicians generally fall on the Pro-Peace side, but they become more polarized after the Indian airstrike and reach a peak following the Pakistani airstrike. This is consistent with reported quotes by Pakistani officials, saying that the Pakistani airstrike was designed to avoid escalation. Similarly, tweets by Indian politicians from the INC and other parties become strongly Pro-Peace directly following the Indian airstrike, with polarity increasing after the Pakistani airstrike. In contrast, on the day of the Pakistani airstrike, tweets by BJP politicians remain Pro-Aggression, possibly focusing either on praise for the Indian airstrike or condemnation of the Pakistani airstrike. The polarity of the BJP tweets belatedly switches to Pro-Peace on the following day (2/28), though the strength of the Pro-Peace polarity still remains weaker for BJP tweets than for tweets by other politicians.

Conclusions While §3.1 and §3.2 present methodology for analyzing a news outlet, in this section, we develop an approach that takes advantage of the short-text nature of Twitter through co-occurrences networks, as well as the strong semantic signals provided by hashtags Ferragina et al. (2015). This method facilitates analyzing how user polarities can change over time in a multilingual corpus, allowing us to show how Twitter users in India and Pakistan used polarizing language during a period of escalating tensions between the two nations. Polarizing language on social media has become an opinion manipulation strategy that can have long-lasting sociopolitical impacts. Our methodology offers tools that require minimal supervision and can be used in multilingual settings to facilitate future work in this area.

3.4 Conclusions and Future Work

This chapter presents several methods and case studies for identifying and characterizing subtle opinion manipulation strategies. Given the subtle nature and global scale of misinformation and propaganda, this line of work focuses on developing methods that can process multilingual text and detect manipulation in the absence of ground-truth labels. There remain numerous directions for exploration in this line of work, including developing methods to detect strategies other than agenda setting, framing, and polarization, discovering finer-grained or corpus-specific frames, and extending methods to other types of media. Beyond this line of work, which focuses on *detection*, additional research is needed to investigate *effectiveness* and *mitigation*.

While news outlets or politicians may employ agenda setting, framing, or polarization strategies, the effectiveness of these strategies in the modern era of online information is unclear. For example, when a news outlet focuses on international news instead of economic downturns, are readers actually distracted from economic downturns? Or do they call out the media for biased coverage? This type of question could be investigated by comparing media coverage with sources of public opinion, including survey data or social media. In general, little research has examined the dynamics between social media and mainstream media (which more effectively achieves agenda-setting and drives public attention?). Time-series analyses like Granger-causality have the potential to provide insight into these dynamics by facilitating examinations of whether topics in social media drive topic coverage of mainstream or the inverse. Some of the challenges of research in this direction include the difficulty in measuring public opinion, particularly given the use censorship as a form of public narrative manipulation (King et al., 2013).

Furthermore, detection serves as a first step towards *mitigation*. Automated methods to uncover propaganda can be used to develop information aggregators that aim to mitigate them, for example, by ensuring the selection of content to aggregate contains a variety of frames. Similarly, aggregation sites and social media platforms could avoid promoting content that is polarizing or promote content from both sides of a poll. Further research is needed to develop these types of mitigation strategies and examine their effectiveness.

Chapter 4

Identifying Toxicity on Social Media

The chapter discusses work previously published in Field and Tsvetkov (2020) and Xia et al. (2020).

The prevalence of toxic content on social media platforms is a widely publicized problem. Given the vast amount of content and the mental toll that manual review has on human moderators (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Silva et al., 2016; Mondal et al., 2017; Mathew et al., 2019), there is need for automated methods, which has prompted much NLP research on detecting hate speech and offensive language e.g. (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018). However, numerous challenges remain. Models trained on hand-labeled data sets are prone to overfitting to shallow surface-level features, resulting in false positives that can dispropriationally affect marginalized populations (Dixon et al., 2018; Davidson et al., 2019; Sap et al., 2019). Additionally, as offensiveness is subjective and context-dependent, annotators of different backgrounds or viewing different instructions assign different offensiveness labels, making it difficult to rely on these judgements (Waseem, 2016; Breitfeller et al., 2019; Sap et al., 2019). Furthermore, most popular data sets have been generated by searching social media platforms like Twitter for posts containing offensive terms from a pre-defined list and then crowd-sourcing toxicity annotations (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018), and models trained on this type of data fail to identify comments that contain subtle or non-explicit toxicity (Breitfeller et al., 2019; Jurgens et al., 2019).

In this chapter, we develop methods to address some of these challenges by drawing frameworks from causal inference and fairness. First, motivated by prior work showing that models trained on hate speech data are liable to label any text containing features of African American English as offensive (Sap et al., 2019; Davidson et al., 2019), we employ an adversarial training approach in order to discourage hate speech classifiers from learning content predictive of dialect correlations (§4.1). In this context, race or dialect is considered a *protected attribute*. Second, we develop methodology to identify systemic differences in social media comments addressed towards men and women (§4.2). This method aims to capture gender bias without relying on hand-annotated data, since bias can be subtle, implicit, and difficult for annotators to recognize. The primary challenge in this work involves structuring the model to learn features predictive of bias, rather than ones that are correlated with gender but are not indicative of harmful content.

Thus, while this chapter focuses addressing challenges in toxic language detection, it more broadly develops methodology to reduce the influence of confounding variables or protected attributes in text classification models. These methods can be used in other settings to encourage model generalizability and reduce overfitting to shallow lexical features. These features are essential in modeling social-oriented tasks, where content is often subtle and dependent on context.

4.1 Demoting Racial Bias in Hate Speech Detection

4.1.1 Background

Most datasets currently used to train hate speech classifiers were collected through crowdsourced annotations (Davidson et al., 2017; Founta et al., 2018), which risks annotator bias. Waseem (2016) show that non-experts are more likely to label text as abusive than expert annotators, and Sap et al. (2019) show how lack of social context in annotation tasks further increases the risk of annotator bias, which can in turn lead to the marginalization of racial minorities. More specifically, annotators are more likely to label comments as abusive if they are written in African American English (AAE). These comments are assumed to be incorrectly labelled, as annotators do not mark them as abusive if they are properly primed with dialect and race information (Sap et al., 2019).

These biases in annotations are absorbed and amplified by automated classifiers. Classifiers trained on biased annotations are more likely to incorrectly label AAE text as abusive than non-AAE text: the false positive rate (FPR) is higher for AAE text, which risks further suppressing an already marginalized community. More formally, the disparity in FPR between groups is a violation of the Equality of Opportunity criterion, a commonly used metric of algorithmic fairness whose violation indicates discrimination (Hardt et al., 2016). According to Sap et al. (2019), the false positive rate for hate speech/abusive language of the AAE dialect can reach as high as 46%.

Thus, Sap et al. (2019) reveal two related issues in the task of hate speech classification: the first is biases in existing annotations, and the second is model tendencies to absorb and even amplify biases from spurious correlations present in datasets (Zhao et al., 2017; Lloyd, 2018). While current datasets can be re-annotated, this process is time-consuming and expensive. Furthermore, even with perfect annotations, current hate speech detection models may still learn and amplify spurious correlations between AAE and abusive language (Zhao et al., 2017; Lloyd, 2018).

In this section, we present an adversarial approach to mitigating the risk of racial bias in hate speech classifiers, even when there might be annotation bias in the underlying training data. In §4.1.2, we describe our methodology in general terms, as it can be useful in any text classification task that seeks to predict a target attribute (here, toxicity) without basing predictions on a protected attribute (here, AAE). Although we aim at preserving the utility of classification models, our primary goal is not to improve the raw performance over predicting the target attribute (hate speech detection), but rather to reduce the influence of the protected attribute.

In §4.1.3 and §4.1.4, we evaluate how well our approach reduces the risk of racial bias in hate speech classification by measuring the FPR of AAE text, i.e., how often the model incorrectly labels AAE text as abusive. We evaluate our methodology using two types of data: (1) a dataset inferred to be AAE using demographic information (Blodgett et al., 2016), and (2) datasets annotated for hate speech (Davidson et al., 2017; Founta et al., 2018) where we automatically infer AAE dialect and then demote indicators of AAE in corresponding hate speech classifiers. Overall, our approach decreases the dialectal information encoded by the hate speech model, leading to a 2.2–3.2 percent reduction in FPR for AAE text, without sacrificing the utility of hate speech classification.

4.1.2 Methodology

Our goal is to train a model that can predict a target attribute (abusive or not abusive language), but that does not base decisions off of confounds in data that result from protected attributes (e.g., AAE dialect). In order to achieve this, we use an adversarial objective, which discourages the model from encoding information about the protected attribute. Adversarial training is widely known for successfully adapting models to learn representations that are invariant to undesired attributes, such as demographics and topics, though they rarely disentangle attributes completely (Pryzant et al., 2017; Li et al., 2018; Elazar and Goldberg, 2018; Lample et al., 2019; Kumar et al., 2019; Landeiro et al., 2019).

Model Architecture Our demotion model consists of three parts: 1) An encoder H that encodes the text into a high dimensional space; 2) A binary classifier C that predicts the target attribute from the input text; 3) An adversary D that predicts the protected attribute from the input text. We used a single-layer bidirectional LSTM encoder with an attention mechanism. Both classifiers are two-layer MLPs with a tanh activation function.

Training Procedure Each data point in our training set is a triplet $\{(x_i, y_i, z_i); i \in 1 \dots N\}$, where x_i is the input text, y_i is the label for the target attribute and z_i is label of the protected attribute. The (x_i, y_i) tuples are used to train the classifier C, and the (x_i, z_i) tuple is used to train the adversary D.

We adapt a two-phase training procedure from Kumar et al. (2019). We use this procedure because Kumar et al. (2019) show that their model is more effective than alternatives in a setting similar to ours, where the lexical indicators of the target and protected attributes are closely connected (e.g., words that are common in non-abusive AAE and are also common in abusive language datasets). In the first phase (pre-training), we use the standard supervised training objective to update encoder H and classifier C:

$$\min_{C,H} \sum_{i=1}^{N} \mathcal{L}(C(H(x_i)), y_i) \tag{4.1}$$

After pre-training, the encoder should encode all relevant information that is useful for predicting the target attribute, including information predictive of the protected attribute.

In the second phase, starting from the best-performing checkpoint in the pre-training phase, we alternate training the adversary D with Equation 4.2 and the other two models (H and C) with Equation 4.3:

$$\min_{D} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(D(H(x_i)), z_i) \tag{4.2}$$

$$\min_{H,C} \frac{1}{N} \sum_{i=1}^{N} \alpha \cdot \mathcal{L}(C(H(x_i)), y_i) + (1-\alpha) \cdot \mathcal{L}(D(H(x_i)), 0.5) \tag{4.3}$$

Unlike Kumar et al. (2019), we introduce a hyper-parameter α, which controls the balance between the two loss terms in Equation 4.3. We find that α is crucial for correctly training the model (we detail this in §4.1.3).

We first train the adversary to predict the protected attribute from the text representations outputted by the encoder. We then train the encoder to "fool" the adversary by generating representations that will cause the adversary to output random guesses, rather than accurate predictions. At the same time, we train the classifier to predict the target attribute from the encoder output.

4.1.3 Dataset

To the best of our knowledge, there are no datasets that are annotated both for toxicity and for AAE dialect. Instead, we use two toxicity datasets and one English dialect dataset that are all from the same domain (Twitter):

DWMW17 (Davidson et al., 2017) A Twitter dataset that contains 25K tweets annotated as *hate speech*, *offensive*, or *none*. The authors define hate speech as language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group, and offensive language as language that contains offensive terms which are not necessarily inappropriate.

FDCL18 (Founta et al., 2018) A Twitter dataset that contains 100K tweets annotated as *hateful*, *abusive*, *spam* or *none*. This labeling scheme was determined by conducting multiple rounds of crowdsourcing to understand how crowdworkers use different labels. Strongly impolite, rude, or hurtful language is considered abusive, and the definition of hate speech is the same as in DWMW17.

BROD16 (Blodgett et al., 2016) A 20K sample out of a 1.15M English tweet corpus that is associated with African American twitter users as determined by a topic model aligned with census demographic information. Further analysis shows that the dataset contains significant linguistic features of African American English.

In order to obtain dialect labels for the DWMW17 and FDCL18, we use an off-the-shelf demographically-aligned ensemble model (Blodgett et al., 2016) which learns a posterior topic distribution (topics corresponding to African American, Hispanic, White and Other) at a user, message, and word level. Blodgett et al. (2016) generate a AAE-aligned corpus comprising tweets from users labelled with at least 80% posterior probability as using AAE-associated terms. Similarly, following Sap et al. (2019), we assign AAE label to tweets with at least 80% posterior probability of containing AAE-associated terms at the message level and consider all other tweets as Non-AAE.

	Accuracy		F1	
	base	ours	base	ours
DWMW17	**91.90**	90.68	75.15	**76.05**
FDCL18	**81.18**	80.27	66.15	**66.80**

Table 4.1: Accuracy and F1 scores for detecting abusive language. F1 values are macro-averaged across all classification categories (e.g. hate, offensive, none for DWMW17). Our model achieves an accuracy and F1 on par with the baseline model.

	Offensive		Hate	
	base	ours	base	ours
FDCL18-AAE	20.94	**17.69**	3.23	**2.60**
BROD16	16.44	**14.29**	5.03	**4.52**

Table 4.2: False positive rates (FPR), indicating how often AAE text is incorrectly classified as hateful or abusive, when training with the FDCL18 dataset. Our model consistently improves FPR for offensiveness, and performs slightly better than the baseline for hate speech detection.

In order to obtain toxicity labels for the BROD16 dataset, we consider all tweets in this dataset to be non-toxic. This is a reasonable assumption since hate speech is relatively rare compared to the large amount of non-abusive language on social media (Founta et al., 2018). We refer to Xia et al. (2020) for details on the model architecture and training parameters.

4.1.4 Results and Analysis

Table 4.1 reports accuracy and F1 scores over the hate speech classification task. Despite the adversarial component in our model, which makes this task more difficult, our model achieves comparable accuracy as the baseline and even improves F1 score. Furthermore, the results of our baseline model are on par with those reported in Sap et al. (2019), which verifies the validity of our implementation.

Next, we assess how well our demotion model reduces the false positive rate in AAE text in two ways: (1) we use our trained hate speech detection model to classify text inferred as AAE in BROD16 dataset, in which we assume there is no hateful or offensive speech and (2) we use our trained hate speech detection model to classify the test partitions of the DWMW17 and FDCL18 datasets, which are annotated for hateful and offensive speech and for which we use an off-the-shelf model to infer dialect, as described in §4.1.3. Thus, for both evaluation criteria, we have or infer AAE labels and toxicity labels, and we can compute how often text inferred as AAE is misclassified as hateful, abusive, or offensive.

Notably, Sap et al. (2019) show that datasets that were annotated for hate speech without sufficient context—like DWMW17 and FDCL18—may suffer from inaccurate annotations, in that annotators are more likely to label non-abusive AAE text as abusive. However, despite the risk of inaccurate annotations, we can still use these datasets to evaluate amplification of racial bias in toxicity detection models. A high FPR over the corresponding test sets suggests that the classification model amplifies bias in the training data, and labels non-toxic AAE text as toxic even when annotators did not. However, given the possibility of annotator bias (Sap et al., 2019), FPRs over the DWMW17 and FDCL18 test sets are likely lower-bounds, and the true FPR is could be even higher.

Table 4.2 reports results for both evaluation criteria when we train the model on the FDCL18

	Offensive		Hate	
	base	ours	base	ours
DWMW17-AAE	**38.27**	42.59	**0.70**	2.06
BROD16	**23.68**	24.34	**0.28**	0.83

Table 4.3: False positive rates (FPR), indicating how often AAE text is incorrectly classified as hateful or offensive, when training with DWMW17 dataset. Our model fails to improve FPR over the baseline, since 97% of AAE-labeled instances in the dataset are also labeled as toxic.

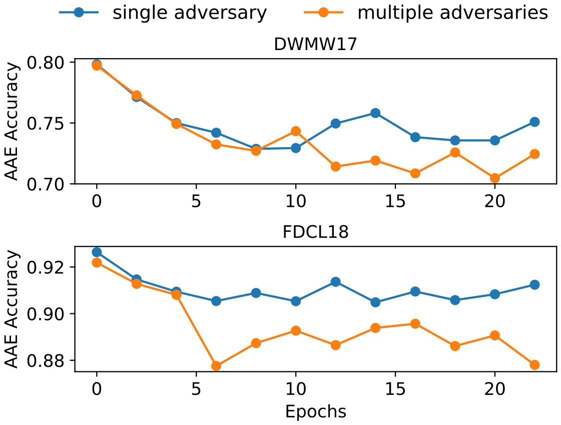


Figure 4.1: Validation accuracy on AAE prediction of the adversary in the whole training process. The green line denotes the training setting of one adversary and the orange line denotes the training setting of multiple adversaries.

data. In both cases, our model successfully reduces FPR. For abusive language detection in the FDCL18 test set, the reduction in FPR is > 3; for hate speech detection, the FPR of our model is also reduced by 0.6 compared to the baseline model. We can also observe a 2.2 and 0.5 reduction in FPR for abusive speech and hate speech respectively when evaluating on BROD16 data.

Table 4.3 reports results when we train the model on the DWMW17 dataset. Unlike Table 4.2, unfortunately, our model fails to reduce the FPR rate for both offensive and hate speech of DWMW17 data. We also notice that our model trained with DWMW17 performs much worse than the model trained with FDCL18 data. In the DWMW17 data, the vast majority of tweets labeled as AAE by the dialect classifier were also annotated as toxic (97%). Thus, the subset of the data over which our model might improve FPR consists of merely $< 3\%$ of the AAE portion of the test set (49 tweets). In comparison, 70.98% of the tweets in the FDCL18 test set that were labeled as AAE were also annotated as toxic. Thus, we hypothesize that the performance of our model over the DWMW17 test set is not a representative estimate of how well our model reduces bias, because the improvable set in DWMW17 is too small.

In Figure 4.1, we plot the validation accuracy of the adversary through the entire training process in order to verify that our model does learn a text representation at least partially free of dialectal information. Further, we compare using one adversary during training with using multiple adversaries (Kumar et al., 2019). Through the course of training, the validation accuracy of AAE prediction decreases by about 6–10 and 2–5 points for both datasets, indicating that dialectal information is gradually removed from the encoded representation. However, after a certain training threshold (6 epochs for DWMW17 and 8 epochs for FDCL18), the accuracy of the classifier (not shown) also drops drastically, indicating that dialectal information cannot be completely eliminated from the text representation without also decreasing the accuracy of hate-speech classification. Multiple adversaries generally cause a greater decrease in AAE prediction than a single adversary, but do not necessarily lead to a lower FPR and a higher classification accuracy. We attribute this to the difference in experimental setups: in our settings, we focus on one attribute to demote, whereas Kumar et al. (2019) had to demote ten latent attributes and thus required multiple adversaries to stabilize the demotion model. Thus, unlike in Kumar et al. (2019), our settings do not require multiple adversaries, and indeed, we do not see improvements from using multiple adversaries.

While we focus on AAE dialect and toxicity in this work, our task set-up and adversarial training approach readily generalize to other settings, such as reducing bias related to age, gender, or income-level in any other text classification task. Overall, our approach has the potential to improve fairness and reduce bias in NLP models.

4.2 Unsupervised Discovery of Implicit Gender Bias

4.2.1 Background

In §4.1, we show that adversarial training can reduce the effects of correlations between a protected attribute and a classification label on a machine-learning classifier. However, the approach in §4.1 focuses on reducing model tendencies to absorb and amplify spurious correlations in data sets; it does not directly address the other challenge in toxic language detection identified by Sap et al. (2019): annotator bias. While Sap et al. (2019) focus on the tendency of annotators to falsely label AAE text as offensive, in general concepts like offensiveness, toxicity, stereotypes, and prejudice are difficult to define and identify, especially for non-experts. In this chapter, we develop an unsupervised approach to detecting implicit gender bias in text that does not rely on subjective human annotations.

In NLP, much literature has examined biases in data, algorithms, or model performance (Sun et al., 2019). This chapter looks further up the pipeline and draws inspiration from the theory that biases in data originate in human cognition. Social biases appear to be a natural component of human cognition that allow people to make judgments efficiently (Kahneman et al., 1982). As a result, they are often *implicit*—people are unaware of their own biases (Blair, 2002; Bargh, 1999)—and manifest subtly, e.g., as microaggressions or condescension (Huckin, 2002; Sue, 2010). These characteristics suggest that biases frequently manifest in text—psychology studies often examine human perceptions through word associations (Greenwald et al., 1998)—but are difficult for human annotators to identify.

Prior work NLP examinations of bias in text conduct broad corpus-level analyses or rely on supervised models. While corpus-level analyses, e.g. associations between gendered words and stereotypes, can be insightful (Bolukbasi et al., 2016; Fast et al., 2016; Caliskan et al., 2017; Garg et al., 2018; Friedman et al., 2019; Chaloner and Maldonado, 2019), they are difficult to interpret over short text spans. They also often rely on human-defined "known" stereotypes, such as lists of traditionally male and female occupations obtained through crowd-sourcing, which restricts analysis to a narrow surface-level domain. Similarly, supervised approaches can provide insight into carefully defined types of bias (Wang and Potts, 2019; Breitfeller et al., 2019; Sap et al., 2020), but they rely

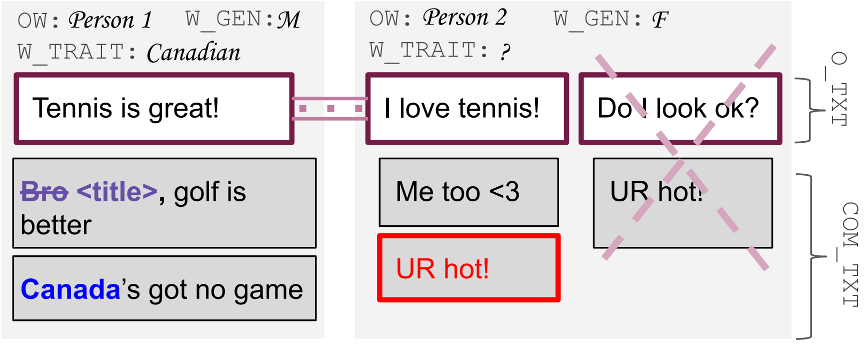


Figure 4.2: We train a classifier to predict the gender (WRITER_GENDER) of the person that text is addressed to (COMMMENT_TEXT), while demoting features that are predictive of gender but not predictive of bias. Posts with similar content are matched through propensity scores; unmatched posts are discarded. Latent traits of the addressee (e.g., nationality) are demoted through an adversarial objective. Overtly gendered language ("Bro") is substituted. Comments indicative of gender despite these restrictions are likely to contain bias.

on human annotations tasks, which are difficult to design or generalize to other domains, especially because social concepts differ across contexts and cultures (Dong et al., 2019).

Our work offers a new approach to surfacing gender bias that does not require direct supervision and is meaningful at a sentence or paragraph level. We create a model that takes text in the 2nd-person perspective as input and predicts the gender of the person the text is addressed to. If the classifier predicts the gender of the addressee with high confidence based only on the text directed to them, we hypothesize that the text is likely to contain bias. The main challenge is encouraging the model to focus on text features that are indicative of bias, rather than artifacts in data that correlate with the gender of the addressee but occur because of confounding variables (*confounds*). Thus, the core of our methodology focuses on reducing the influence of confounds. Our goal is not to improve accuracy of the gender-prediction task, but rather to validate that our methodology demotes confounds and surfaces comments likely to contain gender bias.

In §4.2.2, we define the problem and intuition behind our approach. We describe our methods for confound demotion in §4.2.3, and we evaluate them in §4.2.5. Our evaluation involves examining how confound control affects performance on in-domain and out-of-domain classification tasks, including detection of gender-based microaggressions. Our results suggest that our model successfully identifies text likely to contain bias against women, allowing us to analyze how this bias differs across domains (§4.2.6). To the best of our knowledge, this is the first work that aims to analyze bias in short text spans by learning implicit associations from data sets.

4.2.2 Problem Formulation

Our primary task is to detect gender bias in a communicative domain, specifically in texts targeting an addressee (i.e., 2nd-person) without relying on explicit bias annotations. Our goals align with a causality framework in that we seek to identify content that occurs because of the gender of the addressee rather than other factors. We can define a counterfactual: *Would the addressee have received different text if their gender were different?*

While our framework is broadly applicable, in order to define consistent notation, we consider a setup where our primary text is a comment written in reply to text written by someone else. This includes domains like replies on social media posts, or comments on newspaper articles, and can be generalized other media, e.g., comments on YouTube videos. We identify the following variables:

- ORIGINAL_WRITER: "Original Writer", the person who wrote the original text, e.g., the addressee
- ORIGINAL_TEXT: content of the original text
- WRITER_GENDER: the gender (**M**, **F**) of ORIGINAL_WRITER. We use a binary variable because all of the individuals in our corpus identify as men or women, but our methodology is generalizeable and can be used to examine bias against other genders.
- WRITER_TRAITS: any traits of ORIGINAL_WRITER other than gender, e.g., social role, age, nationality.
- COMMMENT_TEXT: comments replying to ORIGINAL_TEXT

Our goal is to detect bias in COMMMENT_TEXT values that occurs because of WRITER_GENDER. A naive approach would train a classifier to predict WRITER_GENDER from COMMMENT_TEXT and assume that any COMMMENT_TEXT values for which the classifier correctly predicts WRITER_GENDER with high confidence contain bias. However, COMMMENT_TEXT may contain features that are predictive of WRITER_GENDER but are not indicative of bias.

For example, in Figure 4.2, when the comment "UR hot!" (COMMMENT_TEXT) is addressed to someone who said "I love tennis!" (ORIGINAL_TEXT), it is an objectification and unsolicited reference to appearance, which could indicate bias. However, when it is addressed to someone who said "Do I look ok?", it is likely not indicative of bias. If women ask "Do I look ok?" more frequently than men, this naive classifier would identify "UR hot!" is likely addressed towards a woman and identify it as biased. However, we only want the model to learn that references to appearance are indicative of gender if they occur in unsolicited contexts. Thus our model needs to account for the effects of ORIGINAL_TEXT: Because of correlations between WRITER_GENDER and ORIGINAL_TEXT, COMMMENT_TEXT values may contain features that are *predictive* of WRITER_GENDER, but are *caused* by ORIGINAL_TEXT, rather than by WRITER_GENDER. We face a similar problem with WRITER_TRAITS. From the synthetic example in Figure 4.2, if our data set contains more men from Canada than women, the model might learn that references to Canada indicate WRITER_GENDER = **M**. We provide additional empirical examples in §4.2.4.

We refer to factors that might influence COMMMENT_TEXT as *confounding variables* and the artifacts that they produce in COMMMENT_TEXT as *confounds*. We distinguish two types: *observed* and *latent*. Latent confounding variables cannot be controlled if they are entirely unknown; instead, we assume there are observed signals that can be used to infer them, but the values themselves are difficult to explicitly enumerate. In addition to confounds introduced by ORIGINAL_TEXT and WRITER_TRAITS, COMMMENT_TEXT may also contain overt signals, e.g. titles like "Ma'am" or "Sir", that are predictive of gender, but not indicative of bias. We thus identify 3 factors to account for: ORIGINAL_TEXT, WRITER_TRAITS, and overt signals.

4.2.3 Methodology

Our overall methodology centers on creating a classifier that predicts gender of the addressee while controlling for the effects of observed confounding variables (ORIGINAL_TEXT), latent confounding variables (WRITER_TRAITS), and overt signals. The input to the prediction model is COMM-

MENT_TEXT, while the output is WRITER_GENDER, and we aim to identify bias in COMMMENT_TEXT.

Controlling Observed Confounding Variables through Propensity Matching

Our primary method for controlling for ORIGINAL_TEXT is *propensity matching*. Propensity matching was developed to replicate the conditions of randomized trials in causal inference studies (Rosenbaum and Rubin, 1983, 1985). In this step, we discard any COMMMENT_TEXT training samples whose associated ORIGINAL_TEXT is heavily affiliated with only one gender. In Figure 4.2, if we assume that only women post "Do I look ok?", we would discard all comments posted in reply to the ORIGINAL_TEXT "Do I look ok?". We ultimately seek to balance our data set, so that the set of all COMMMENT_TEXT where WRITER_GENDER = **M** has similar associated ORIGINAL_TEXT as the set of all COMMMENT_TEXT where WRITER_GENDER = **F**. Thus, we match each ORIGINAL_TEXT where WRITER_GENDER = **F** with a similar ORIGINAL_TEXT where WRITER_GENDER = **M** and discard all unmatched data.

Ideally, we would match ORIGINAL_TEXT values written by men with identical ORIGINAL_TEXT values written by women, but this is infeasible in practice. Instead, the key insight behind propensity matching is that it is sufficient to match data points based on the probability of the target variable, e.g., the probability that WRITER_GENDER = **F** (Rosenbaum and Rubin, 1983, 1985). Thus, the propensity score e_i for a COMMMENT_TEXT$_i$ is defined as the probability that WRITER_GENDER = **F**, given the confounding variable, ORIGINAL_TEXT$_i$:

$$e_i(\text{COMMMENT\_TEXT}_i) = P(\text{WRITER\_GENDER}_i = \mathbf{F} | \text{ORIGINAL\_TEXT}_i)$$

To balance our data set, we need to ensure that the set of COMMMENT_TEXT where WRITER_GENDER = **M** has a similar propensity score distribution as the set of COMMMENT_TEXT where WRITER_GENDER = **F**. Because propensity scores are dependent on ORIGINAL_TEXT, all COMMMENT_TEXT replied to the same ORIGINAL_TEXT have the same propensity score. We can then equate $e_i(\text{COMMMENT\_TEXT}_i) = e_i(\text{ORIGINAL\_TEXT}_i)$, and focus estimating ORIGINAL_TEXT scores.

Propensity scores can be estimated by using a classification model that is trained to predict the target attribute WRITER_GENDER$_i$ = **F** from the observed confounding variable ORIGINAL_TEXT$_i$ (Westreich D, 2010; Lee et al., 2010). We use a bidirectional LSTM encoder followed by two feed-forward layers with a tanh activation function and a softmax in the final layer. Then, we use greedy matching to match each ORIGINAL_TEXT$_i$ where the true value of WRITER_GENDER$_i$ is **F** with ORIGINAL_TEXT$_j$ where the true value of WRITER_GENDER$_j$ is **M** and $|e_i(\text{ORIGINAL\_TEXT}_i) - e_j(\text{ORIGINAL\_TEXT}_j)|$ is minimal (Gu and Rosenbaum, 1993).

We institute a threshold c (Stuart, 2010), where we discard ORIGINAL_TEXT$_i$ if we cannot find a ORIGINAL_TEXT$_j$ such that $|e_i(\text{ORIGINAL\_TEXT}_i) - e_j(\text{ORIGINAL\_TEXT}_j)| \leq c$. Thus, for example, we would match a post written by a woman that is "stereotypically female" (e.g., e_i is large) with a post written by a man that is also "stereotypically female" (e.g., e_j is also large). In Figure 4.2, we match "Tennis is great" with "I love tennis", and we discard "Do I look ok?" as unable to be matched. However, using propensity matching rather than direct matching allows us to match ORIGINAL_TEXT values that are about different topics, as long as they are equally likely to have been written by a woman.

Finally, our actual model input consists of COMMMENT_TEXT, not of ORIGINAL_TEXT. Once we have matched pairs of ORIGINAL_TEXT values, we need to ensure that we have an equal number of COMMMENT_TEXT values for each ORIGINAL_TEXT in the pair in order to have a balanced data set. Then, for each matched [ORIGINAL_TEXT$_i$, ORIGINAL_TEXT$_j$], we randomly downsample to have an equal number of COMMMENT_TEXT values for each ORIGINAL_TEXT in the pair. In this way, we balance the training set of COMMMENT_TEXT in terms of how predictive the confounding variable ORIGINAL_TEXT is of the target attribute WRITER_GENDER.

Controlling Latent Confounding Variables through Adversarial Training

While propensity matching is a desirable way to control for confounding variables because of established literature, matching is only possible for observed variables (Gu and Rosenbaum, 1993; Rosenbaum, 1988). In our data, while ORIGINAL_TEXT is observed, WRITER_TRAITS is not possible to match on (further discussion in §4.2.4). Instead, we use an adversarial objective drawn from Kumar et al. (2019) to encourage the model to ignore WRITER_TRAITS.

Confound representation While we cannot explicitly enumerate WRITER_TRAITS, we know that they are associated with the identity of ORIGINAL_WRITER, and we can infer them from COMMMENT_TEXT addressed to ORIGINAL_WRITER. We use associations between ORIGINAL_WRITER and COMMMENT_TEXT to derive a feature vector for each COMMMENT_TEXT$_i$ that reflects WRITER_TRAITS$_i$. The latent confounds to demote are represented as multinomial distributions, derived from log-odds scores (Monroe et al., 2008).

For each label ORIGINAL_WRITER $= k$ and each word type w in all COMMMENT_TEXT, we calculate the log-odds score $lo(w, k) \in \mathbf{R}$, where higher scores indicate stronger associations between k and the word. In Figure 4.2, $lo(\texttt{Canada}, \texttt{Person 1})$ would be high, as COMMMENT_TEXT values addressed to `Person 1` often contain the word Canada. Then, following Kumar et al. (2019), we define a distribution: for all $k \in$ ORIGINAL_WRITER and an input COMMMENT_TEXT$_i$, $= \langle w_1, \ldots, w_n \rangle$:

$$p(k|\text{COMMMENT\_TEXT}_i) \propto$$
$$p(k)p(\text{COMMMENT\_TEXT}_i|k) = p(k)\prod_{i=1}^{n} p(w_i|k)$$

$p(k)$ is estimated from the distribution of k in the training data, i.e., the proportion of COMMMENT_TEXT values addressed to ORIGINAL_WRITER $= k$. $p(w_i|k)$ is proportional to $\sigma(lo(w, k))$, where we use the sigmoid function (σ) to map log-odds scores to the range [0,1] and then normalize them over the vocabulary to obtain valid probabilities. For each input COMMMENT_TEXT$_i$, we then obtain a vector whose elements are $p(k|\text{COMMMENT\_TEXT}_i)$ and whose dimensionality is the number of ORIGINAL_WRITER individuals in the training set. We normalize these vectors to obtain multinomial probability distributions which reflect COMMMENT_TEXT$_i$'s association with each ORIGINAL_WRITER individual. Thus, when we demote this vector during training, we force the classifier to learn features that are indicative of the group WRITER_GENDER and not features that are indicative of individual members of this group (e.g., some group members are from Canada). We refer to the confound vector as t_i. Justification for the log-odds representation as opposed to alternatives is presented in Kumar et al. (2019).

Training Procedure Our goal is to obtain a model that can predict WRITER_GENDER, but cannot predict the latent confounds represented by t_i. To achieve this, the model is trained in an alternate GAN-like procedure (Goodfellow et al., 2014).

First, the input $x \in$ COMMMENT_TEXT is encoded using an encoder neural network $h(x;\theta_h)$ to obtain a hidden representation $\mathbf{h}_x$. This representation is then passed through two feedforward networks: (1) $c(h(x);\theta_c)$ to predict the label $y \in \{\mathbf{M}, \mathbf{F}\}$; and (2) an adversary network $\text{adv}(h(x);\theta_a)$ to predict the vector representation of the latent confounds.

We train the encoder, so that $\mathbf{h}_x$ does not contain any information predictive of the confound vector, but does contain information predictive of the target attribute. Thus our primary training objective is:

$$\min_{c,h} \frac{1}{N} \sum_{i=1}^{N} \text{CE}(c(h_{x_i}), y_i) + \text{KL}(\mathbb{U}_K \| \text{adv}(h_{x_i}))$$

where $\mathbb{U}$ represents a uniform distribution, CE represents cross-entropy loss, and KL represents KL-divergence. We refer to Kumar et al. (2019) for the training procedure that alternates minimizing this objective and training the adversary.

Overt Signals

Finally, we control for overt signals using word substitutions that replace gendered terms with more neutral language, for example woman → ⟨person⟩ and man → ⟨person⟩. We create a 66-term list of substitutions from existing resources (Zhao et al., 2018; Bolukbasi et al., 2016) as well as our observations of the data. We also use substitutions to remove the names of addressees from comment, replacing ORIGINAL_WRITER's "Firstname" and "Lastname" with "⟨name⟩" in COMMMENT_TEXT. We do not attempt to identify nicknames, as the confound demotion method described in §4.2.3 should already mitigate the influence of individual names, and we perform the substitution as merely an extra precaution.

4.2.4 Experimental Setup

Our primary data is the Facebook subsection of the RtGender corpus (Voigt et al., 2018). The data contains two subsections: *Politicians* (400K posts and 13.9M replies addressed to 412 then-current U.S. members of Congress), and *Public Figures* (118K posts and 10.7M replies addressed to 105 famous people such as actresses and tennis players).

We can show that ORIGINAL_TEXT is a confounding variable by computing log-odds scores between the words in ORIGINAL_TEXT and WRITER_GENDER (Monroe et al., 2008). In the Politicians data, the most female-associated words are *women*, *Congresswoman*, *sexual*, and *assault*. The most male-associated words are *Obamacare*, *Iran*, *EPA*, and *spending*. It is evident that male and female politicians post about different topics, e.g., female politicians likely post more about sexual assault. A naive model may predict that comments using sexual language are addressed towards women, but increased sexual language may occur because of ORIGINAL_TEXT, rather than gender bias.

A similar problem occurs with WRITER_TRAITS, e.g., the corpus has more comments addressed to female tennis players (9 players; 184K comments) than male players (1 player; 29K comments). The model can obtain high accuracy by predicting WRITER_GENDER = **F** for COMMMENT_TEXT with the word "tennis". Unlike ORIGINAL_TEXT, which is observable from the data, we have no way

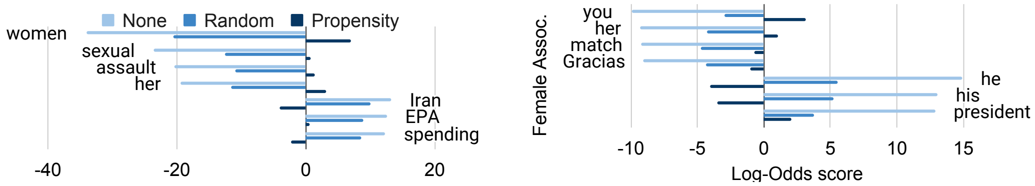


Figure 4.3: Log-odds scores for most polar words in Politicians (left) and Public Figures (right) data, with different matching methods. Propensity matching best reduces polarity.

of enumerating every possible value in writer_traits. Even if we could enumerate them, we do not expect propensity matching over writer_traits to work, because we cannot find reasonable matches, e.g., there is only one senior senator from Massachusetts. Additionally, writer_traits can be as fine-grained as names: we cannot find a male senator whom commenters call "Liz Warren".

We divide each data set into train, dev, and test sets, enforcing no original_writer overlap between subsets. We perform propensity matching and derive the confound vectors to demote using only the training data. We apply word substitutions to all subsets.

4.2.5 Evaluation

We train our model to predict writer_gender from commment_text, employing propensity matching over original_text, word substitutions over commment_text, and writer_traits demotion. We focus on evaluating how well our model controls for confounds and whether or not it captures gendered language. Successful demotion of confounds would suggest that our model learns to identify text indicative of gender bias.

Observed Confounding Variable Demotion In Figure 4.3, we show log-odds scores, measuring association between original_text and writer_gender in the training set before and after propensity matching. For comparison, we also show scores for a randomly matched data set, in which we balance original_text to have an equal proportion of **F** and **M** labels by random sampling (constructed to be the same size as the propensity matched set). In the Politicians and Public Figures data, propensity matching reduces the magnitude of the most polar words: log-odds scores for the matched data are closer to zero than for the non-matched or randomly matched data.[1] Further, propensity matching can even cause the polarity to change direction: words that were originally female-associated (e.g. "her") become slightly male-associated. These figures suggest that propensity matching effectively reduces the confounding influence of original_text.

Latent Confounding Variable Demotion We evaluate how well our model demotes the influence of latent confounding variables over the held-out test sets (Table 4.4). We created data splits so that there is no overlap in original_writer values between the train and test sets. While there may still be overlap in some latent writer_traits, we expect there to be less overlap in writer_traits between the train and test set than within the train set. Thus, improved performance over the held-out test set would suggest that demotion effectively reduces the influence of the latent confounding

[1] Polarities were reduced without producing new ones: in the Politicians data, the magnitude of the 2 most polar words decreased from -34.0 and 17.9 to -7.68 and 8.52, and in the Public Figures data, from -45.5 and 39.3 to -5.29 and 9.43.

	Public Figs		Politicians	
	F1	Acc.	F1	Acc.
base	74.9	63.8	23.2	73.2
+demotion	**76.1**	**65.1**	17.4	**77.1**
+match	65.4	56.0	28.5	46.7
+match+dem.	68.2	59.7	**28.8**	51.4

Table 4.4: Evaluation over held-out test sets, where WRITER_GENDER = **F** is considered the positive class. Latent confound demotion improves performance.

variables—the model learns characteristics of comments addressed to women generally rather than characteristics specific to the individual people in the training set. We do not necessarily expect propensity matching to improve performance, as this method reduces the influence of confounding variables that have high overlap between the train and test sets.

Because the data set is imbalanced (the Politicians test set is 82%**M** and the Public Figures test set is 35.9%**M**), we report F1 and accuracy scores in Table 4.4, where WRITER_GENDER = **F** is considered the positive class. As expected, models with demotion perform best on all metrics, with the exception of recall in the Politicians data. We note that in general lack of performance improvement on the test set does not necessarily mean the model is not working, and it could indicate that there is not biased language in the data set. However in this case, since we do observe biased comments in this data (e.g. Table 4.6), and we do observe a performance increase, the performance increase suggests that confound demotion improves the model's ability to generalize beyond the individuals in the training set and capture characteristics of language addressed to women in general.

Detection of Sexist Comments Finally, we evaluate if our model captures gender-biased language by using it to identify gender-based microaggressions, i.e.,"you're too pretty to be a computer scientist!". This task is notoriously difficult because words like "pretty" often register as positive content (Breitfeller et al., 2019; Jurgens et al., 2019). Our goal is not to maximize accuracy over microaggression classification, but rather to assess whether or not our model has encoded any indicators of gender bias from the RtGender data set, which would be indicated by better than random performance.

We use a corpus of self-reported microaggressions taken from Breitfeller et al. (2019), who collected the corpus from www.microaggressions.com. On this website, posters describe a microaggression that they experienced. They can using quotes, transcripts, or narrative text to describe the experience, and these posts are tagged with type of bias expressed, such as "gender", "ableism", "race", etc. We discard all posts that contain only narrative text, since it is not 2$^{\text{nd}}$ person perspective and thus very different than our training data, which leaves 1,604 posts for analysis. In the absence of negative examples that contain no microaggressions, we focus on distinguishing gender-tagged microaggressions (704 posts) from other forms of microaggressions, e.g., racism-tagged (900 posts). We train our model on either the Politicians or Public Figures training data sets, and then we test our model on the microaggressions data set. Because most gender-related microaggressions target women, if our model predicts that the reported microaggression was addressed to a woman (e.g. WRITER_GENDER = **F**), we assume that the post is a gender-tagged microaggression. Thus, our models are *not trained at all* for identifying gender-tagged microaggressions.

	Public Figs		Politicians	
	F1	Acc.	F1	Acc.
base	61.3	57.3	48.1	64.2
+demotion	**62.2**	57.9	53.7	61.5
+match	38.9	55.9	46.9	50.7
+match+dem.	50.9	57.0	**56.9**	49.9
Random	46.0	49.8	-	-
Class Random	42.1	48.3	-	-

Table 4.5: Evaluation over the microaggressions data set. Despite not being trained for this task, our models achieve better-than-random performance.

Table 4.5 shows results from our models and two random baselines. "Random" guesses gender-tagged or not with equal probability. "Class Random" guesses gender-tagged or not according to true test distributions (56.1% gender-tagged). All models outperform "Class random", and all models with demotion also outperform "Random".

Propensity matching improves F1 when training on the Politicians data, but not Public Figures. Several differences could explain this: the Public Figures set is smaller, so propensity matching causes a more substantial size reduction. Also, the Politicians data is more heavily imbalanced, though notably, it is imbalanced in the same direction as the microaggressions data, while the Public Figures data is imbalanced oppositely. Finally, many microaggressions contain references to appearance, which are also common in the Public Figures data. Many comments to people like actresses focus on their looks, especially because they often post photos. However, by controlling for ORIGINAL_TEXT, propensity matching discards many of these comments. Thus, by demoting a confounding variable, we make the prediction task more difficult. Our goal in confound demotion is not to improve accuracy, but to increase confidence in model outputs.

Nevertheless, the general better-than-random performance of all models is striking, as it suggests strong bias in the underlying training data, which is encoded by our models.

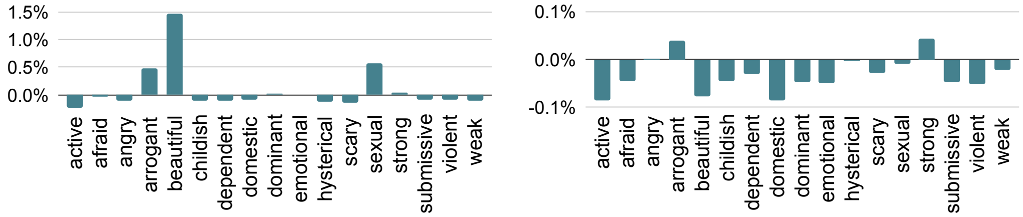


Figure 4.4: Lexicon differentials between comments with a high likelihood of bias and random samples with WRITER_GENDER = **W** for Public Figures (left) and Politicians (right) data. In the Public Figures data, high-likelihood comments are more focused on appearance.

4.2.6 Analysis of Encoded Bias

Finally, we analyze what type of bias our model learns: (1) we identify words that most impact model confidence; (2) we compare posts surfaced by our model with prior work on stereotypes; (3) we show example posts surfaced by our model. Throughout this section, we use *prediction score* to refer to the output of the final softmax layer of the prediction model, which we take as an

Politicians	
ORIGINAL_TEXT	From reintroducing my legislation to curb sexual assault on college campuses to...
COMMMENT_TEXT	DINO I hope another real Democrat challenges you next election
ORIGINAL_TEXT	Donald Trump is the President, not our ruler...Speak up! Call the White House...
COMMMENT_TEXT	⟨name⟩ Shea-Porter, I did not vote for you and have no clue why anyone should have. You do not belong in politics
Public Figures	
ORIGINAL_TEXT	I am wondering about the guy who actually cried over spilt milk? He must have had...
COMMMENT_TEXT	Total tangent I know but, you're gorgeous.
ORIGINAL_TEXT	Bob and I join Bill Hemmer on America's Newsroom to discuss whether or not...
COMMMENT_TEXT	I like Bob, but you're hot, so kick ⟨theirs⟩ butt.

Table 4.6: Example comments surfaced by our model from Politicians and Public Figures data sets.

estimate of model confidence. We generally focus on COMMMENT_TEXT for which our model predicts WRITER_GENDER = **F** with a high prediction score. These are the posts our model identifies as likely to contain bias against women: despite the matching and demotion methods, the model still predicts WRITER_GENDER = **F** with high confidence.

Influential words We identify words that strongly influence the model's decisions by masking out words from comments in the test set and examining the impact on prediction score. For each data set, we take the 500 comments from the test set for which the model predicts WRITER_GENDER = **F** with maximal prediction scores. We then generate masked posts: for every word w in the post, we generate a version of the post that omits w. We run these masked posts through our gender-prediction model and compare the prediction scores where w is omitted and where w is not omitted, averaging across all occurrences of w in the 500 posts. We then examine the set of w words with the highest differential in prediction score - these are words that, when omitted, cause the model to less associate WRITER_GENDER with **F**.

In the Public Figures data, the most influential words are appearance-driven and sexualized: *beautiful*, *bellissima*, *amore*, *amo*, *love*, *linda*, *sexo*. In contrast, influential words in the Politicians data are more mixed. Words include references to strength and competence, e.g., *force*, *situation*, as well as traditionally domestic terms, e.g., ⟨*spouse*⟩[2], *family*, *love*. When we repeat this process using the 500 highest-confidence posts from the training set instead of the test set, we find similar results. Influential words in the Public Figures training data primarily refer to appearance, while ones in the Politicians training data include terms like *DINO*.[3] However, influential words from the training data also includes some correlative terms, like names of states, that we would expect the latent confound demotion to de-emphasize. While §4.2.5 suggests that our model successfully reduces the influence of confounding variables, more work is needed to eliminate them entirely.

Comparison to stereotype lexicons In order to better understand these trends, we draw from prior work on stereotype detection (Fast et al., 2016). We take the set of test comments for which our model predicts WRITER_GENDER = **F** with a high prediction score (≥ 0.99 for Public Figures; ≥ 0.95 for Politicians). Then, we compute the difference in frequency of words from a stereotype lexicon (Fast et al., 2016) in this high-confidence prediction set with their frequency in a random

[2] "⟨⟩" indicate overt terms substituted out. "⟨spouse⟩" replaced "husband", "husbands", "wife", and "wives".
[3] "Democrat in Name Only" a political insult

sample of the same number of comments where the true value of WRITER_GENDER = **F**.[4]

Figure 4.4 reports results, which reflect the same trends observed in the influential words. In the Public Figures data, the lexicons that overlap the most with the high-bias posts are "Beautiful", "Arrogant", and "Sexual", which suggests that bias in these comments focuses on appearance and sexualization. In contrast, bias in comments directed towards politicians are less focused, and differences between the high-confidence prediction posts and the random sample are smaller. The two most prominent lexicons are "Arrogant" (primarily driven by lexicon words *special*, *proud*) and "Strong". Notably, we do not account for negation of lexicon words. A narrative of power is reflected in comments surfaced by our model: "you & Nikki Haley lost my vote on the flag issue *your both weak*". We provide more examples in Table 4.6.

Because the stereotype lexicons are small and scores can be dominated by a few words, we also compare LIWC scores (Pennebaker et al., 2001). While most LIWC categories are too broad to align with well-known stereotypes, results are consistent with Figure 4.4; for Public Figures, the high-bias data scores higher than the random sample for the "Sexual" (0.32 vs. 0.10) and "Body" (0.70 vs. 0.56). For Politicians, the high-bias comments score lower than the random sample in the "Drives" (8.76 vs. 9.71), which encompasses Affiliation, Achievement, Power, Reward, and Risk focus.

The difficulty in evaluating our model against existing lexicons as well as the differences between the two data sets motivates our goal in learning to detect bias automatically. Bias can differ in different contexts, making it difficult to crowdsource through annotations or define through lexicons.

Examples Table 4.6 shows training and test examples surfaced by our model. We identify them by selecting posts where ORIGINAL_TEXT is not strongly gendered (propensity score model described in §4.2.3 outputs a prediction score < 0.6), but where COMMMENT_TEXT is strongly gendered (> 0.9 prediction score). While posts from the Politicians data are diverse, posts from the Public Figures data focus on appearance. These comments reflect the broader trends shown in the influential words and in Figure 4.4.

4.3 Conclusions and Future Work

Detection of content that is offensive or biased is useful for fostering civil communication on social media as well as in other settings, like workplace communications. Tools to identify potentially harmful content can be used to remove it, avoid promoting it, and encourage users to revise their statements, as bias is often implicit and unintentional. While this chapter offers initial approaches toward combating some of the challenges in toxic language detection, including algorithmic unfairness, annotator bias, and content subtlety, we identify several limitations and areas for future work.

Both §4.1 and §4.2 use adversarial training to demote the influence of confounding variables on a classification model. Results from both sections suggest that adversarial training does help reduce the influence of confounding variables, but it does not eliminate them completely and there is scope for improvement. The goal of reducing the influence of data confounds has alignment with

[4] We ignore non-English comments and lemmatize the comment text and lexicons. We randomly sample twice and average frequencies between samples. Lexicon counts are normalized by total number of words in the sample.

causal inference, and the intersection of causal inference and NLP is a growing area of research (Chandrasekharan et al., 2017; Roberts et al., 2020; Egami et al., 2018; Veitch et al., 2020; Keith et al., 2020).

Additionally, the evidence suggesting that human judgements are not reliable for detecting toxic or biased comments makes evaluation difficult. §4.1 likely under-reports model bias and §4.2 focuses on evaluating independent parts of the pipeline and uses an external evaluation task. Research in this area would benefit from the development of additional evaluation metrics and data sets, such as annotated data that includes information about preceding context and relevant social variables and modeling approaches that gracefully handle annotator disagreement (Sap et al., 2020).

There are also directions for future work within the frameworks established in this chapter. For example, in §4.2 while we focus on some confounds in the data, there may be additional ones that our model does not account for, such as the impact of videos, photos, or links shared with ORIGINAL_TEXT. Processing this data requires the integration of multimodal techniques. Similarly, while our model uses ORIGINAL_TEXT for propensity matching in the training data, thus encouraging the model to encode indicators of bias, a model to classify comments as biased or unbiased should also incorporate ORIGINAL_TEXT when assessing test data. Additionally, we focus on the perspective of ORIGINAL_WRITER and examine what bias social media users may be exposed to, i.e. what comments men and women might expect to receive in response to their posts. In this work, we do not examine why comments addressed toward men and women may differ, whether because the same commenters write different comments to men and women, or because men and women attract comments from different types of people. In §4.1 we do take more of commenter-centric framing, but we focus only on one variable: commenter's use of AAE. Overall, detecting and analyzing bias is a first step towards mitigating it, and we hope our work will encourage future work in this area.

Chapter 5

Supporting Policy Decisions Through Text Analysis

Much of the work presented in this thesis has relevance to policy decisions. For example, understanding information manipulation strategies can inform regulatory legislation and aid technology companies in determining how to prioritize mitigation (§3). However, this chapter focuses more explicitly on work whose primary goal is providing information to decision-makers and organizers in evolving and sensitive social settings.

In the first part (§5.1), we analyze a data set of tweets related to Black Lives Matter protests from May 24th to June 28th, 2020 using a domain adaptation model for measuring emotions perceived in tweets about specific events. In the past few decades, social psychologists have recognized the important role emotions play in activism: "moral shocks" can facilitate people joining a movement, while hope and pride are necessary to sustain involvement (Jasper, 1997; Goodwin et al., 2007; Jasper, 2011; Ma, 2017). Understanding the dynamics between emotions (such as what balance between anger and optimism produces a "hopeful anticipation of impact" that motivates continued action) can both provide insight into past movements and guidance for future efforts (Jasper, 2011; Goodwin et al., 2007; Allen and Leach, 2018).

Furthermore, projected emotions have been used to falsely characterize Black people, leading to tangible harms. For example, the "angry Black woman" stereotype can result in negative physical, social and economic impacts, such as facilitating workplace discrimination (Collins, 1990; Walley-Jean, 2009). In the context of social movements, negative stereotypes of Black protesters as violent angry "thugs" have long been used to derail civil rights activism (Leopold and Bell, 2017).[1] Analyzing emotions in tweets about protests can provide evidence refuting these types of negative portrayals.

In the second part (§5.2), we examine the opportunities and pitfalls of integrating text features into the risk predictive tool used by the Allegheny County Department of Human Services in processing referrals about child welfare. Child welfare cases involve copious amounts of notes written by caseworkers and service providers, which contain professional assessments of the family situation, risks, and needs (Saxena et al., 2020; DePanfilis, 2003). However, these notes are often too

[1] https://www.nbcnews.com/news/us-news/not-accident-false-thug-narratives-have-long-been-used-discredit-n1240509

numerous for caseworkers and supervisors to review manually when making time-sensitive decisions or examining a case they are unfamiliar with (Perron et al., 2019). Developing NLP models to incorporate text features in existing predictive models could improve model performance, leverage data that is currently not referenced manually, and ultimately help child welfare agencies improve the services they offer. The broader potential for NLP to aid in processing expert-written notes has been demonstrated in other domains, including healthcare and law (Uzuner et al., 2011; Johnson et al., 2016; Ji et al., 2021; Zhong et al., 2020).

However, the incorporation of free-form text could also exacerbate many of the concerns around risk assessment and algorithmic decision-making: text is written by people and reflects their perceptions of events, which may or may not accurately reflect reality (Eberhardt, 2020). Model tendency to absorb and amplify data biases can lead to unfairness in downstream tasks (Bolukbasi et al., 2016; Zhao et al., 2017; Nangia et al., 2020; Zhang et al., 2020; Amir et al., 2021), though much remains unknown about how these effects manifest in high-stakes decision settings (Blodgett et al., 2020; Field et al., 2021). These concerns are particularly pronounced in a setting like child welfare where families involved in child welfare services already express mistrust in "the system" (Brown et al., 2019) and there is pronounced criticism of using algorithmic tools in any capacity (Eubanks, 2018).

§5.2 seeks to further understanding the benefits and risks of using natural language in predictive risk assessments. This work has the potential to impact policy decisions at multiple levels, including the decision of whether or not to incorporate natural language features in algorithmic tools in high-stakes settings, as well as the decisions made in individual child welfare cases if these tools are implemented. Additionally, these questions are relevant more generally to the debate on how to resolve concerns around the use of algorithms in high-stakes decision making. Recent calls for disentangling the role of technical versus sociotechnical interventions to resolve biases in algorithmic risk assessments come amidst a broader debate over when to repair or abolish the use of algorithms in high stakes settings (Kluttz et al., 2018; Roberts, 2019; Benjamin, 2019; Selbst et al., 2019). We aim to probe the limits of one possible technical fix: augmenting the data features with natural language. In this sense, our work is part of the tradition of computing as *rebuttal* (Abebe et al., 2020).

Thus, this chapter explores the potential impact of language processing on informing organizers and decision makers in addressing social issues. We highlight some of the potential benefits, including debunking stereotypes about Black Lives Matter protesters, as well as some of the risks, such as the potential to increase racial disparity in the child welfare system

5.1 Analysis of Emotions in #BlackLivesMatter Tweets

5.1.1 Background

The term *#BlackLivesMatter* originated in posts made by activists Alicia Garza and Patrisse Cullors in 2013, following police officer George Zimmerman's acquittal over the killing of Trayvon Martin, an unarmed Black teenager (Richardson, 2019).[2] The term has since become popularized as referring to movements against police brutality and the extrajudicial killing of Black people. These movements

[2] https://blacklivesmatter.com/about/ We generally use "Black Lives Matter" to refer to the broad movement against police brutality, rather than the official organization.

have continually grown and evolved, garnering widespread attention following the deaths of Michael Brown in Ferguson and Eric Garner in New York (2014) (Bonilla and Rosa, 2015; Choudhury et al., 2016), and more recently, George Floyd in Minneapolis (2020). The death of George Floyd, in addition to the deaths of Ahmaud Arbery and Breonna Taylor, led to widespread protests against police violence and racism.[3]

Social media has been an integral part of these movements. In addition to *#BlackLivesMatter*, millions of tweets were posted with hashtags like *#Ferguson*, *#JusticeForGeorgeFloyd*, and *#ICantBreathe*. While forms of "digital protest" and "hashtag activism" can occur organically, they are often a tool used by community activists, who may plan hashtag campaigns, promote in-person activism, and intentionally bypass traditional media (Bonilla and Rosa, 2015; Freelon et al., 2016; Richardson, 2019; Choi et al., 2020; Jackson et al., 2020). Thus, social media not only provides an avenue for analyzing modern social movements but understanding social media messaging is also essential for providing insight into these events.

As discussed in §5, understanding emotions expressed on social media is particularly useful for building effective social movements and debunking stereotypes. However, measuring emotions is non-trivial, and computational models that overestimate expressions of emotions like "anger" can reinforce negative stereotypes. Previous examinations of emotions and affect expressed in tweets about the Black Lives Matter movement have relied on lexicon (LIWC) scores (Choudhury et al., 2016), and analyses of other protest events have similarly relied on lexicon-based approaches (Mohammad and Turney, 2013; Steinert-Threlkeld and Joo, 2020). While recent research has led to the development of more powerful deep-learning based models and annotated data sets, these models nevertheless are prone to over-fitting to shallow lexical cues and often perform poorly in new domains (Mohammad et al., 2018; Demszky et al., 2020; Desai et al., 2020). Thus, in this work, we leverage recent natural language processing (NLP) techniques, including in-domain pre-training and few-shot learning, to improve emotion analysis model performance across new domains, in an easily-adaptable framework. We evaluate our model using two annotated data sets of emotion classification in two different domains, Reddit and Twitter, and for six different emotion categories: ANGER, DISGUST, POSITIVITY, SURPRISE, FEAR, and SADNESS (Ekman, 1992).

We ultimately use our model to examine emotion trends in a data set we collected containing $\sim$ 34 million tweets related to the Black Lives Matter movement. In examining estimated perceived emotions over time, in tweets with specific hashtags, and in comparison to on-the-ground protests, our results consistently identify the prominence of positivity (e.g., pride, optimism, excitement), which supports social theories about the importance of emotions like hope and pride and offers evidence countering "angry Black" stereotypes.

5.1.2 Methodology

Data Our primary corpus consists of tweets about Black Lives Matter. We gathered English tweets posted between May 24th and June 30th, 2020 using the Twitter search API. §A.1.1 contains the full list of terms used for data collection, which includes terms likely to be used by both supporters and critics of the Black Lives Matter protests. Our final data set, which we refer to as #BLM2020, consists of 250M tweets (34.7M, excluding retweets) by 18.9M users. Figure 5.1 presents the volume

[3] https://acleddata.com/2020/09/03/demonstrations-political-violence-in-america-new-data-for-summer-2020/

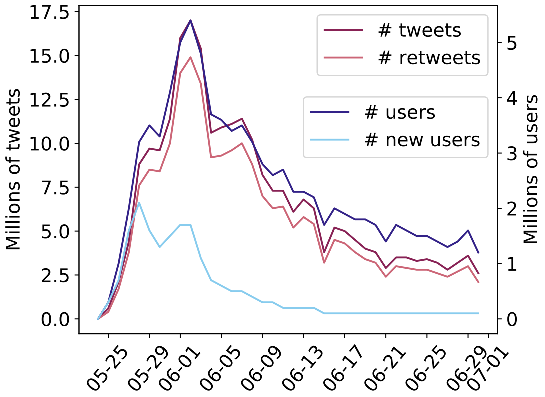


Figure 5.1: Distribution of tweets, retweets, users, and new users in #BLM2020.

of tweets and users through the time span. There is high Twitter engagement in the first 10 days, followed by a slow decrease in the subsequent four weeks. We ceased data collection at the end of June, given the substantial decline in tweet volume by the end of the month.

Detecting emotions expressed in tweets In order to analyze emotions expressed in #BLM2020, we develop and evaluate models for identifying 6 emotion categories: ANGER, DISGUST, FEAR, POSITIVITY, SURPRISE, and SADNESS, which are the primary core emotions according to Ekman's taxonomy (Ekman, 1992). This approach assumes emotions identified by annotators in tweets can be represented in discrete categories, and taking a psychological constructionist perspective of measuring emotions, e.g. focusing on the dimensions of valence/sentiment and arousal/agency as in §2 may have different results (Barrett and Russell, 2014). We follow prior work in considering these 6 Ekman emotions to be supersets of finer-grained emotions (Demszky et al., 2020):

- ANGER: anger, annoyance, disapproval, rage
- DISGUST: disgust, loathing, boredom
- FEAR: fear, nervousness, vigilance, apprehension
- POSITIVITY[4]: amusement, approval, excitement, gratitude, love, optimism, relief, pride, admiration, desire, caring, acceptance, anticipation, serenity, trust, ecstasy
- SURPRISE: realization, confusion, curiosity, amazement, distraction
- SADNESS: disappointment, embarrassment, grief, remorse, pensiveness

Throughout this work, we treat emotions as non-exclusive (e.g., a tweet may contain both ANGER and SADNESS). We also aim to capture emotions that Twitter users choose to express or solicit on the platform, which may not reflect their actual emotional state, and we discuss this distinction in our analysis.

Traditionally, social scientists have used lexicon-based approaches to measure emotions in tweets, determining whether or not a tweet expresses "anger" based on whether or not it contains any words from a list of "angry" words. While lexicon-based approaches remain popular because of their ease-of-use they can be brittle and fail to adapt to new domains. Word connotations change in

[4] Ekman's taxonomy refers to this dimension as "joy", but defines it as encompassing all positive emotions (Ekman, 1992). We use the term "positivity" instead of "joy" throughout this work to reflect the breath of emotions encompassed by this dimension and avoid suggesting that happiness is a dominant emotion in #BLM2020

different contexts (Field et al., 2019), particularly in protest movements, which often aim to subvert the status quo. For example, EmoLex, a large NRC emotion lexicon of word associations with 8 emotions, associates *police* with *fear*, *positive*, and *trust*, which is contradictory to the connotations of *police* in protests against police brutality (Mohammad and Turney, 2010, 2013). More recently, machine-learning based NLP models have outperformed traditional lexicon approaches at identifying affect in text (Demszky et al., 2020; Desai et al., 2020; Rasooli et al., 2018; Potts et al., 2021). Neural models are trained on annotated datasets and used to infer affect in unseen text. However, a model trained on a pre-collected data set may still perform poorly on data from a different domain where connotations differ. Collecting new annotated datasets for every domain of interest is prohibitively time-consuming and expensive, especially for tasks that require in-domain knowledge or involve subjective judgements.

Instead, we take a *domain adaptation* approach: given a set of *source* data annotated for perceived emotional content (for example, tweets with binary present/not present labels for emotions like ANGER, SURPRISE, and FEAR), our goal is to infer emotion labels for a set of *target* data from a different domain using explicit methods to adapt the model to this new domain. Different domains could include text about a different event or from a different social media platform. Domain adaptation allows us to re-use annotated data sets, rather than collecting new annotated data for every domain of interest. We train and evaluate a base classifier for inferring emotions with two variants of domain adaption:

Base classifier (BASE) In the simplest setting, we train a prediction model over the annotated source data and infer labels on the target data without any explicit domain adaptation. We specifically use a pre-trained language model (BERT) fine-tuned over the source data (details in §A.1.2).

Task-adaptive pre-training (+TGT) NLP has recently seen large performance improvements through masked language model pre-training: models are pre-trained by optimizing them to predict words that have been obfuscated from input sentences (Devlin et al., 2019). The same model can then be fine-tuned for a specific task. Following prior work, we use masked language model pre-training over unannotated sentences from the target data to encourage domain adaption, and then fine-tune the model to infer emotions using the annotated source data, as in BASE (Howard and Ruder, 2018; Gururangan et al., 2020; Desai et al., 2020).

Few-shot learning (+FSL) While collecting a large annotated data set for every new domain can be infeasible, collecting annotations over a small number of in-domain labeled data is often practical. In this model, we fine-tune the classifier over small sets of annotated target data (300 instances), after training over the larger source data set.

Our primary training data is drawn from two sources, GoEmotions and HurricaneEmo (Demszky et al., 2020; Desai et al., 2020). GoEmotions consists of 58,000 English Reddit comments manually labeled for emotion categories or neutral (Demszky et al., 2020). We randomly divide this data into train (80%), validation (10%), and test (10%) splits. HurricaneEmo consists of 15,000 English tweets about hurricanes annotated for 24 emotions according to Plutchik's scheme (Desai et al., 2020; Plutchik, 2001), which we map to the Ekman scheme. The original data set provided a different

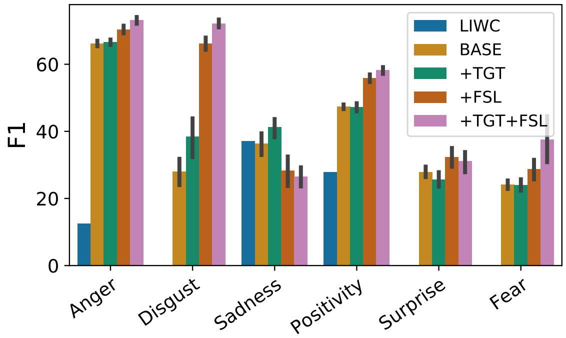


Figure 5.2: F1 scores of emotion classifiers evaluated over #BLM2020. Error bars indicate the 95% CI.

train-test split for each emotion, thus we created our own instance-level data split of train (70%), validation (10%), and test (20%). To facilitate few-shot learning and evaluation, we additionally collect emotion annotations over 700 randomly-sampled tweets from #BLM2020 using the 6 Ekman emotions, and we use 300 as training data, 100 as development data, and 300 as test data. We provide further details in §A.1.2.

Figure 5.2 shows evaluation results over the annotated #BLM2020 test data, where we use both GoEmotions and HurricaneEmo as training data, and use 300 of the annotated #BLM2020 for few-shot learning. We provide additional validation metrics over larger test data sets in Appendix A.1.3. In addition to classification models, we provide LIWC as a baseline, since it is a popular dictionary-based analysis method and has previously been used in analyzing tweets about Black Lives Matter (Tausczik and Pennebaker, 2010; Choudhury et al., 2016). We map the LIWC dimensions of "Anger", "Positive Emotion", and "Sadness" to ANGER, POSITIVITY, and SADNESS, respectively, since they are the only emotions that directly map to LIWC dimensions, and we map the floating-point scores produced by LIWC to binary labels using the best-performing threshold over the validation data set.

The machine learning classifiers generally outperform LIWC, few-shot learning brings a large performance improvement, and +TGT+FSL achieves the best overall performance. As +TGT+FSL outperforms other models, we use it to obtain perceived emotion labels for all tweets in #BLM2020, which we analyze in the following section. We generally focus our analysis on the emotions that our model identified with highest F1 and that had the highest inter-annotator agreement in our annotated data (reported in §A.1.2): ANGER, DISGUST, and POSITIVITY. Performance of the +TGT+FSL model is poor for SADNESS in Figure 5.2, but SADNESS is very sparse in the #BLM2020 test set, and +TGT+FSL outperforms other models when evaluated over a larger test set (§A.1.3). In contrast, SURPRISE has poorer model performance and lower inter-annotator agreement over all test sets. Thus, we avoid extended discussion of this emotion, though we do display metrics for all emotions.

5.1.3 Analysis of Emotions in #BLM2020

We first use our inferred emotion labels to examine how emotions expressed in #BLM2020 change over time and with different hashtags. Because retweets are not written independently nor displayed as separate posts to Twitter users, and we do not expect model performance to be reliable over very short tweets, we exclude retweets and tweets with < 5 tokens, leaving 34.1M tweets for analysis.

Changes in emotions over time In Figure 5.3a, we plot the percentage of tweets that contain each emotion over time estimated using our model. Although POSITIVITY captures a broader range of emotions than ANGER, ANGER is the most prevalent emotion throughout, consistently occurring in $> 40\%$ of tweets. POSITIVITY and DISGUST are also prevalent, with POSITIVITY gradually decreasing over time, while ANGER and DISGUST gradually increase after an initial peak. A small peak in SADNESS occurs early on, but is quickly eclipsed. A peak in FEAR occurs on Sunday, May 31 - Monday, June 1, directly following the first weekend of protests (Figure 5.5).

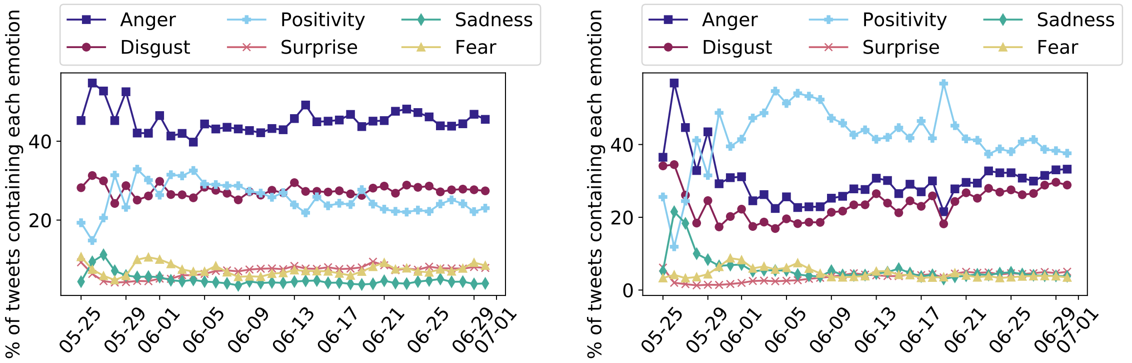
ANGER and POSITIVITY have a strong negative correlation over time (-0.79), while ANGER and DISGUST have a strong positive one (0.69). We also note that annotators who labelled emotions in #BLM2020 described ANGER and DISGUST as difficult to distinguish in this setting (§A.1.2), which is consistent with the identification of "moral outrage" as involving anger and disgust (Salerno and Peter-Hagene, 2013).

Figure 5.3a presents emotions over the entire data set, which contains tweets both supportive of the Black Lives Matter movement and opposed to it. Thus, it provides no insight into how emotions are directed and does not distinguish between, for example, protesters' ANGER and ANGER at protesters. In Figure 5.3b, we display emotion levels only for tweets that contain a pro-BLM hashtag (defined in §A.1.1). Over a subset of $\sim$ 600 tweets that annotators labelled for stance (§A.1.2), using these hashtags to recover tweets annotated as "pro-BLM" obtained a precision of 82.7% and a recall of 29.4%. In Figure 5.3b, the initial peak in SADNESS is even more apparent, as is the high peak of ANGER, both of which pre-date the first weekend protests. POSITIVITY rises shortly before the first weekend and continues through the second weekend before declining. A later peak in POSITIVITY occurs on 06/19/2020, which is Juneteeth, a holiday celebrating the emancipation of people who had been enslaved in the U.S. On this day, #Juneteenth was the second-most common hashtag in the total data set, after #BlackLivesMatter.

Common hashtags for each emotion In Table 5.1, we report hashtags that are most over-represented in tweets our model identifies as containing each emotion, calculated using log-odds with a Dirichlet prior (Monroe et al., 2008). These hashtags are highly indicative of the predicted emotions. Tweets labeled with POSITIVITY commonly contain #love and #pride; tweets labeled with SADNESS contain #sad and #RIP. Importantly, hashtags associated with the same emotions often reflect opposing viewpoints: tweets labeled with ANGER frequently contain both #MAGA (Donald Trump's campaign slogan) and #TrumpResignNow.

Emotions by keywords Figure 5.4 shows the percent of tweets our model identifies as containing each emotion, where tweets are divided as containing pro-BLM hashtags, anti-BLM hastags, terms related to police, and terms related to protests as enumerated in §A.1.1.[5] In all cases, POSITIVITY, ANGER, and DISGUST occur much more than FEAR, SADNESS, and SURPRISE. Both POSITIVITY and SADNESS occur more often in tweets with pro-BLM hashtags than in any of the other subsets. Notably, ANGER and DISGUST are lower in tweets with explicitly pro-BLM hashtags than in tweets with explicitly anti-BLM hashtags, while POSITIVITY is higher. As users often use hashtags to engage

[5] Precision and recall for using anti-BLM hashtags to recover anti-BLM stances are 62.5%/4.5%, and we caution that these results reflect the use of these hashtags, not necessarily stance.



(a) % of all 34.1M tweets

(b) % of 6.5M tweets with pro-BLM hashtags

Figure 5.3: % of tweets containing each emotion over time (May 24th and June 30th). Emotion categories are drawn from Ekman's taxonomy (Ekman, 1992) and inferred over a data set of 34.1M tweets using a neural classification model with domain-adaptation components (+TGT+FSL).

Anger	Disgust	Positivity	Surprise	Sadness	Fear
BreonnaTaylor	Trump	BlackLivesMatter	BLM	RIPGeorgeFloyd	NYCScannerDuty
GeorgeFloyd	Racist	RaiseTheDegree	AllLivesMatter	JusticeForGeorgeFloyd	NYCProtests
DefundThePolice	MAGA	Love	WhiteLivesMatter	GeorgeFloyd	PDX911
PoliceBrutality	TrumpResignNow	BlackOutTuesday	AskingForAFriend	ICantBreathe	COVID19
ACAB	AllLivesMatter	Juneteenth	AlmostBrokeMyHeartAtTheEnd (Thai)	BlackLivesMatter	BlackLivesMatterNYC
Riots2020	BunkerBoy	PrideMonth	Confused	RIP	NYCProtest
GeorgeFloydWasMurdered	DefundThePolice	MatchAMillion	BlackOutTuesday	Sad	DCProtest
MinneapolisRiots	Trump2020	Music	Nkurunziza	JusticeForFloyd	DCProtests
JusticeForGeorgeFloyd	RacistInChief	Art	달빛보다_찬란한_준휘야_생일축하해	RestInPower	Breaking
DerekChauvin	DemocratsAreDestroyingAmerica	Pride2020	HNGInternship	WeAreTired	GeorgeFloydProtests
Trump	AntifaTerrorists	Juneteenth2020	Dollar (Arabic)	RIPHumanity	FoxNews
BreonnaTayor	Democrats	2MforBLM	AmUnbroken	PalestinianLivesMatter	Coronavirus
FakeNews	BLM	Pride	달빛보다_찬란한_준휘	BlackLivesMatters	SeattleProtest
TrumpResignNow	TrumpIsARacist	NYCScannerDuty	Kalu	ShootATweet	NYCScanner
DemocratsAreDestroyingAmerica	ACAB	Equality	365DNI	JusticeForGeorge	Protests
AntifaTerrorists	Antifa	Peace	BlueLivesMatter	JusticeForJeyarajAndFenix	NYPD

Table 5.1: Most common hashtags for tweets labeled each emotion, computed using log-odds with a Dirichlet prior (Monroe et al., 2008). Emotion categories are drawn from Ekman's taxonomy (Ekman, 1992) and inferred using a neural classification model with domain-adaptation components (+TGT+FSL). Hashtags are de-duplicated after case-normalization.

in public narratives and direct content to particular streams (Huang et al., 2010), these data offer counter-evidence to the narrative of BLM protesters as angry "thugs": there is more POSITIVITY and less ANGER and DISGUST in tweets with pro-BLM hashtags (i.e. that are explicitly directed towards streams about the movement) than in tweets discussing these events more generally, including tweets with reactionary #AllLivesMatter hashtags. The highest percentage of ANGER occurs in tweets mentioning police, which encompasses both anger over police brutality and calls for reform, as well as reactionary pro-police posts expressing anger at protesters. The highest percentage of FEAR occurs in tweets mentioning protests, which captures direct references to events that occurred during protests, like aggressive police responses.

Correlations between emotions in tweets and on-the-ground protests Finally, we compare tweet volume and emotions with the volume of on-the-ground protests during the same time period. To estimate on-the-ground protests, we use data collected from two sources: The Armed Conflict Location & Event Data Project (ACLED)[6] (Raleigh et al., 2010) and the Crowd Counting

[6] https://acleddata.com/special-projects/us-crisis-monitor/

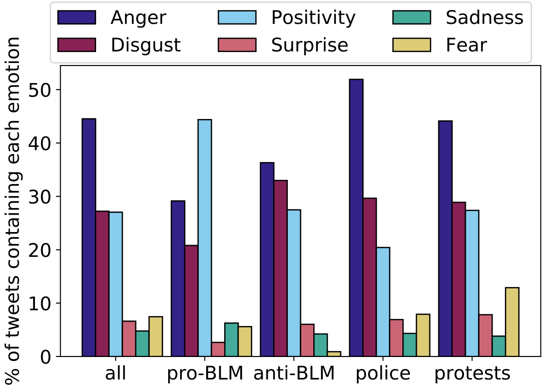


Figure 5.4: Percent of tweets that contain each emotion, where tweets are divided by keywords and hashtags. Emotion categories are drawn from Ekman's taxonomy (Ekman, 1992) and inferred using a neural classification model with domain-adaptation components (+TGT+FSL).

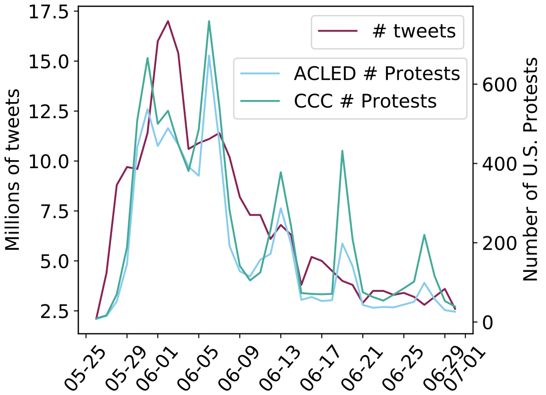


Figure 5.5: Volume of U.S. protests and collected tweets. Protest data is drawn from The Armed Conflict Location & Event Data Project (ACLED) (Raleigh et al., 2010) and the Crowd Counting Consortium (CCC). Twitter data is collected in this work.

Consortium (CCC)[7] (Crowd Counting Consortium, 2022). ACLED contains records of political violence, demonstrations, and strategic developments across the United States. Entries are hand-coded by ACLED researchers and based on media reports by 2,400 sources. CCC contains records of political crowds reported in the United States, including marches, protests, strikes, demonstrations, riots, and other actions and is maintained by a dedicated project manager and research assistants.

Figure 5.5 shows the number of protests across the United States per day, as reported by the ACLED and CCC. Data from both sources show similar patterns, though CCC consistently reports slightly more protest events than ACLED. The first peak in protests occurs on 5/30/2020 - 5/31/2020, the weekend directly following George Floyd's death. The highest peak in Twitter activity occurs after this weekend, which may suggest how early protests called attention to George Floyd's death. The peak volume of protests occurs after the highest peak in Twitter activity, on 06/06/2020, the 2nd Saturday. While definite conclusions cannot be drawn from these few data

[7] https://sites.google.com/view/crowdcountingconsortium/home

	State-level				City-level			
	CCC	p-val	ACLED	p-val	CCC	p-val	ACLED	p-val
ANGER	-0.38	0.0072	-0.42	0.0020	-0.22	0.0001	-0.28	0.0000
DISGUST	-0.19	0.1869	-0.30	0.0356	-0.21	0.0001	-0.27	0.0000
POSITIVITY	0.48	0.0004	0.48	0.0003	0.23	0.0000	0.26	0.0000
SURPRISE	-0.31	0.0267	-0.18	0.2009	-0.04	0.4534	-0.09	0.1009
FEAR	-0.02	0.8867	0.11	0.4517	0.16	0.0040	0.18	0.0012
SADNESS	-0.21	0.1435	-0.40	0.0040	-0.27	0.0000	-0.18	0.0009

Table 5.2: Pearson correlations between % of tweets with each emotion and number of protests in each state or city. Tweets are associated with U.S. cities and states based on locations listed by users in their profiles, where 20.66% of users were aligned to states and 12.3% of users were aligned to cities. Emotion categories are drawn from Ekman's taxonomy (Ekman, 1992) and inferred using a neural classification model with domain-adaptation components (+TGT+FSL). Protest data is drawn from two initiatives: ACLED and CCC.

points, this pattern suggests a possible symbiotic relationship between online and offline protests: the first peak of in-person protests encouraged increased engagement on Twitter, which in turn resulted in even more protests the following weekend. After this weekend, the volume of protests steadily declines, with regular peaks on subsequent weekends.

While protests broke out across the U.S., they were more widespread and long-lasting in certain areas than others, which allows us to compare emotions expressed on Twitter and on-the-ground protests by comparing tweets by users in different locations. We identified location for users in our data set based on the user-populated location string in their profiles (details in §A.1.4). This value was non-empty for 62.36%, of users in our data set, and we were able to map 20.66% of users to a U.S. state and 12.3% of users to U.S. cities listed in the ACELD data.[8] Our results in this analysis are limited to users who specified locations in their Twitter profiles, and we cannot conclude how well they generalize to users who did not, though prior work has suggested that geolocated tweets provide accurate measures of protest events, even though geolocation data is typically sparse (Sobolev et al., 2020; Hecht et al., 2011; Alex et al., 2016).

In Table 5.2 we show the Pearson correlations between the number of protests in each city or state and the percent of tweets containing each emotion as measured by our model, for tweets posted by users in those locations. Because we can expect larger and more populous states to have more protests, we normalize the number of reported protests in each state by the number of counties in the state (a U.S. administrative/political/geographic subdivision of a state with some level of governmental authority), obtaining county counts from U.S. census data reported by SafeGraph.[9] We believe counties are a reasonable normalization term because they reflect factors that influence protests, which typically take place in a single geographic area and are often targeted toward local government. At a city-level, where we do not expect as substantial geographic barriers and given the importance of size in social movements (Chenoweth and Stephan, 2011; Biggs, 2018), we weight protest events by ACLED and CCC size estimates to compute total protest volume (see §A.1.4 for details and discussion).

At both the city and state level, POSITIVITY is positively correlated with more protest events, and ANGER, DISGUST, and SADNESS are negatively correlated. FEAR is additionally positively correlated

[8] Given the sparsity of city-level data, we only compute results for cities for which we were able to identify at least 500 users. We do not observe substantial differences in results if we change the cut-off to 100 or 1000 users

[9] https://www.safegraph.com/

at a city level. Importantly, our results demonstrate *geographic* correlations, not *temporal* ones. We cannot distinguish if expressions of POSITIVITY proceed protests and are thus predictive of on-the-ground activism or if they are reactionary (posted during or after a protest). Given the potential misuses of technology for predicting protests (actors have sought to discourage collective action, de-stabilize movements, or promote polarization (King et al., 2017; Arif et al., 2018)) as well as the non-linear temporal trends in our data (Figure 5.5, protest volume declines over time with spikes on weekends), we do not compute temporal trends or make any attempt to predict protest or tweet volume.

Discussion Political and social psychology research has identified anger as a politically motivating emotion using survey data, laboratory experiments, and theoretical analyses in protest movements and political involvement generally (Valentino et al., 2009, 2011; Pearlman, 2013) as well as specifically for Black people (Banks et al., 2019; Burge, 2020; Scott and Collins, 2020). Our results do not directly contradict this research, in that we do find ANGER and DISGUST as the most commonly expressed emotions in our data set and we see initial peaks in these emotions (Figure 5.3a). However, we find that these emotions are negatively correlated with in-person protests (Table 5.2), whereas POSITIVITY is positively correlated. This difference in result from prior work could result from differences between actual emotional state and what users choose to post on Twitter. Negative stigma around "angry Black" people could disincentive people from posting expressions of anger or disgust on Twitter (Phoenix, 2019). Additionally, as we focus on geographic rather than temporal relations, our model captures emotions expressed before, during, and after protests, and feelings of camaraderie and pride resulting from protests could outweigh other emotions expressed on Twitter.

Relatedly, our results also show positive correlations between FEAR and protests at a city level, which seemingly contradicts prior identification of fear, anxiety, and sadness as dispiriting emotions that deter political engagement (Huddy et al., 2007; Goodwin et al., 2007; Jasper, 2011). However, both SADNESS and FEAR are uncommon in our data, which is consistent with these theories, as posting on Twitter is itself an act of engagement and people feeling sadness or fear may choose not to post at all. An examination of the relatively small percent of tweets that our model does identify as reflecting fear suggests that they often focus on events specifically related to protests, including community monitoring of police activity like severe crowd control tactics during protests (references to "scannerduty" in Table 5.1, higher prevalence of FEAR in tweets referring to protests in Figure 5.4). These results are consistent with the discussion of fear in (Pearlman, 2013)'s analysis of Arab Uprisings, which notes that protesters express fear and suggests identifying conditions under which people press on despite fear is more relevant than identifying conditions under which fear disappears. Unlike FEAR, SADNESS is negatively correlated with protest activity, which is consistent with prior identification of this emotion as dispiriting (Huddy et al., 2007; Valentino et al., 2009; Pearlman, 2013).

Overall, our results consistently identify the role of positive emotions in Black Lives Matter social media posts. In addition to the correlations with on-the-ground protests, tweets with pro-BLM hashtags contain more POSITIVITY than other tweets in our data set, such as ones with anti-BLM hashtags. These results support social psychology theories suggesting that positive emotions are an important component of social movements (Jasper, 2011; Goodwin et al., 2007). While outrage and anger can encourage people to become involved, participants must also have optimism and hope for change, or they will not have the motivation to act (Goodwin et al., 2007; Allen and Leach, 2018). Similarly, joy and camaraderie, e.g., feeling affective bonds as a member of a group, encourage sustained involvement (Goodwin et al., 2007; Jasper, 2011). Our findings additionally also offer evidence countering the narrative of protesters as perpetuating anger.

Prior work on the Black Lives Matter movement has also examined emotions. One study uses LIWC lexicons to measure several dimensions, including *positive/negative affect*, *anger*, *anxiety*, *sadness*, and *swear* in a data set of tweets about Black Lives Matter protests in 2014-2015 (Choudhury et al., 2016). The authors find that *anger* tends to decrease over time, while *friends* and *social* tend to increase, supporting the theory that anger and outrage may cause initial participation, but joy and camaraderie facilitate sustained involvement. They also find that high negativity/sadness but low anger/anxiety on Twitter are predictive of an increased volume of future protests. Beyond language and emotion in Black Lives Matter tweets, other work has examined the motivations and identities of individuals involved, including the prominence of female activists (Richardson, 2019), the demographics of Twitter users (Olteanu et al., 2015), the roles that activists take (Choi et al., 2020), communication networks and widely-shared content (Freelon et al., 2016), estimations of violence using images (Won et al., 2017), and the broader implications of social media activism (Jackson et al., 2020).

While our analysis focuses on tweets about Black Lives Matter, our methodology can be used in other settings, requiring only a small annotated set of in-domain data for fine-tuning and evaluation. These analyses and methodologies can enhance understanding of social movements, providing information to social scientists and activists.

5.2 Opportunities and Perils of Using NLP in Child Welfare

5.2.1 Background

In high-stakes settings such as child welfare, hiring, education, and criminal justice, practitioners are increasingly turning to risk assessment tools to aid humans making time-sensitive high-stakes decisions (Saxena et al., 2020; Vaithianathan et al., 2017; Raghavan et al., 2020; Chouldechova, 2017; Cattell et al., 2021). These risk assessments are often criticized for failing to account for relevant individual context and for automating biases in the data (Whittaker et al., 2018; Eubanks, 2018; Roberts, 2019). Many of these systems rely primarily on tabular structured data, and there is increasing interest in exploring the potential of unstructured data like free-form text for improving these systems by providing contextual information (Saxena et al., 2020; Hsu et al., 2020).

We conduct a case study evaluating the benefits and pitfalls of incorporating features derived from natural language into a pre-existing tool to predict risk of adverse outcomes. First, drawing from prior work on algorithmic tools (Chouldechova et al., 2018; Brown et al., 2019; Fogliato et al., 2021; Jacobs and Wallach, 2021) and the potential for data biases in free-form text notes (Bolukbasi et al., 2016; Zhao et al., 2017; Nangia et al., 2020), we posit several research questions focused on potential opportunities and pitfalls, and we describe evidence-based evaluation methods to investigate them. We then construct several NLP models, including bag-of-words, neural, and pre-trained language models, that incorporate text features into existing predictive tools and evaluate their performance (§5.2.2). Finally, we use these models to empirically investigate the proposed risks and opportunities

(§5.2.3).

Our results show that text features have a small effect on aggregate model performance, but that notes can provide valuable context in specific cases. Similarly, while we do not find evidence of increased algorithmic unfairness in aggregate metrics, we document different patterns of language use for text notes about black versus white families. Additionally, we find that pre-trained models generalize poorly to domain-specific terminology and notes are highly reflective of already-made or imminent decisions. Although we focus on investigating predictive risk models, our findings generally deepen our understanding of the content of notes and therefore have implications for other types of NLP systems trained on this data, such as information extraction or summarization, and even for decisions informed by manual review of notes. Our work serves as a first investigation of free-form text data associated with child welfare cases and provides research questions and experimental frameworks relevant to other domains that involve algorithmic processing of structured data and expert notes.

Overview of child welfare system Child welfare cases typically begin with a referral, where someone (the "reporter") contacts social services with concerns about a child. A call-screening staff member then makes a *Call Screen Decision*: whether or not to investigate the allegations made in the referral. If the referral is *screened in*, a caseworker then conducts an investigation, which may involve interviewing relevant contacts and conducting assessments. In many states, caseworkers must complete the investigation in a fixed amount of time, such as 60 days in Pennsylvania (Bureau, 2017). Based on the investigation, the caseworker decides whether or not the family should be accepted for services by the child welfare agency (*Service Decision*). If the family is accepted, a case is opened. Cases can stay open for varying lengths of time, and families may receive a range of services, such as housing support or addiction treatment often through external *service providers*.

Child welfare agencies increasingly are using predictive risk assessment tools to inform decisions such as which cases to screen in for investigation (Chouldechova et al., 2018; Nash, 2017; Saxena et al., 2020). In Allegheny County, Pennsylvania, the Department of Human Services currently uses a predictive risk assessment tool called the Allegheny Family Screening Tool (AFST) (Vaithianathan et al., 2017). For an incoming referral, the AFST presents call-screening staff with a score from 1 to 20 that aims to reflect the likelihood that the child will be placed (removed from home) within 2 years conditional on the referral being screened in (Chouldechova et al., 2018; Vaithianathan et al., 2017). The model is based on a logistic LASSO that selected 71 features from over 800 variables providing demographics, past welfare interaction, public welfare, county prison, juvenile probation, and behavioral health information on all persons associated with each referral. Some of these structured features are derived from previous interactions with the child welfare system, (e.g. the number of previous referrals associated with people on the new current referral).

In addition to this structured information, previous interactions with the Department of Human Services often result in unstructured free-form text data. In total, as of 2021, Allegheny County has over 3 million *contact notes* written by caseworkers, supervisors, and service providers who have had contact with families. Notes can be written throughout the duration of a case, including during the investigation phase before a case is actually opened, and they are typically associated with a referral, a case, or both. §A.2.1 presents some overview statistics of the full data set of contact notes. While most contact notes record telephone calls or face-to-face contacts, such as visiting families at home or at school, notes can be created for other forms of contact, such as emails. Notes from previous interactions with the Department of Human Services may contain useful information, but features from this data are not currently included in the AFST model. More broadly, in a survey of current algorithms deployed in the U.S. child welfare system, Saxena et al. (2020) report that none of the algorithms surveyed use unstructured text data. Additionally, call screen workers who view risk assessments typically need to make quick decisions on referrals and do not refer to relevant contact notes.

Potential Opportunities In general, natural language from expert assessments offers the promise of uncovering important risk signals and protective (risk-mitigating) factors. In domains like healthcare, incorporating expert (medical) notes into predictive models can improve the model's ability to characterize patient risk (Hsu et al., 2020). In the child welfare setting, we may expect contact notes to offer an in-depth, expert assessment of the family's risk and needs (Eastman et al., 2019). We may posit that information in these notes would be valuable inputs for a predictive model of risk. Saxena et al. (2020) contend that the use of contact notes could improve existing predictive risk assessments by reducing data gaps and incorporating the perspectives of caseworkers. Our work responds to their call for research into this space. We will investigate whether incorporating contact notes uncovers new risk signals not captured in structured data.

Research Question 1 *Do contact notes contain new risk signals not captured in structured data?*

Evaluation: Our evaluation will compare a model built with structured features, which we call "structured only", to a "hybrid" model built with both structured features and contact notes (See §5.2.2 for details on implementation). If contact notes contain new risk signals not captured in structured data, we expect the hybrid model to make fewer mistakes in failing to identify high-risk families relative to the structured only model. In other words, we expect the hybrid model to have a lower false negative rate (FNR) than the structured only model.

Contact notes may additionally be useful in contextualizing risk signals. A growing movement of activists and researchers, many in the FAccT community, argues that often data does not tell the fully story because it fails to account for context, including cultural differences, societal factors like racism and privilege, as well as salient individual circumstances (O'neil, 2016; Eubanks, 2018; Richardson et al., 2019). Stakeholders in the child welfare system such as affected families have expressed concern with algorithmic risk assessments because of the inability to contextualize: "A computer cannot understand context. My son has autism — how does the data account for this?" (Brown et al., 2019). Among other stakeholder concerns was the "deficits-based" nature of the algorithmic risk assessment, that is, that the algorithm attended to risk factors and ignored protective factors that mitigate risk (Brown et al., 2019; Nash, 2017). Consider for instance a family where no father is involved in caregiving. The structured data would summarize this parenting situation as simply single-parent. While in many circumstances this may be a risk factor, a contact note may clarify that a grandmother who lives next-door watches the children while their mother is working. We seek to understand whether contact notes capture these protective factors.

Research Question 2 *Do contact notes provide important context for risk signals or identify pertinent protective factors?*

Evaluation: If contact notes contain protective factors that help the predictive model discern low-risk families, we expect the hybrid model to have a lower false positive rate (FPR) than the structured only model. Our evaluation will also include a qualitative analysis of the text notes for cases whose risk scores notably differ across the hybrid and structured models.

Despite the potential upside to incorporating text features, there are significant potential drawbacks that we consider in the next section.

Potential Pitfalls One of the goals of algorithmic risk assessment systems is to improve the consistency of the decision making process and show that these decisions were unbiased and evidenced-based (Saxena et al., 2020). However, contact notes are not unbiased objective documents. In some instances, caseworkers write notes explicitly to document and justify their decisions, so these notes may be more indicative of the caseworker's decision-making process than of objective risk factors: notes written by a caseworker who expects an out-of-home placement to become necessary and perhaps even eventually initiates placement may reflect this expectation, regardless of whether or not placement was justified by the circumstances of the case. Even when caseworkers intend to represent reality as accurately as possible, cognitive biases may affect what information caseworkers record. Literature on confirmation bias, anchoring, and effort reduction suggests that information that aligns with one's beliefs is more available for recall than conflicting information (Strack and Mussweiler, 1997; Nickerson, 1998; Shah and Oppenheimer, 2008). Then, inputting rich information that is highly reflective of a person's decision into a prediction models may reinforce the model's replication of human decisions, rather than improve consistency or objectivity.

A further complicating factor is that risk assessment models are often trained to predict a proxy outcome because it is impossible to directly observe "ground truth" outcomes, such as harm. The AFST is trained on out-of-home placement, which is both a decision (e.g., a caseworker decides to seek court authorization for placement (O'Connell, 2016)), and also only one possible adverse outcome. As text data is both subjective and high-dimensional, models that incorporate text data are liable to increase over-fitting to these proxy tasks without necessarily improving true estimates of risk.

Research Question 3 *Are contact notes more informative of the author's perspective than of true risk? Does incorporation of contact notes over-fit to proxy tasks?*

Evaluation We investigate this question in two ways. First, we examine how predictive contact notes are of nearby decisions, e.g. how predictive notes from the investigation of a referral are of whether or not the referral is accepted for services. High accuracy on this task suggests that notes are highly reflective of caseworker decision making. Second, we compare how well a model trained to predict out-of-home placements is able to predict other adverse outcomes, similar to the approach in Coston et al. (2020) and De-Arteaga et al. (2021). Adverse events that may signal the child is at risk of harm or neglect include the child experiencing homelessness, a mental health crisis, or involvement with criminal justice. We would generally refrain from using these outcomes as the target of predictive risk assessments for various reasons, including missing data and concerns around discouraging clients from seeking important services like mental health. We nonetheless would hope that the notion of risk learned by training on another outcome would partially transfer to this evaluation task. As our transfer outcome, we use a combined indicator that encompasses six other outcomes including clients receiving mental health support, housing services, or appearing as defendants in court cases (details are provided in Table A.6). For both the nearby decision and transfer tasks, we compare the performances of models with (hybrid) and without (structured) text features.

Finally, stereotypes and prejudice frequently manifest in text (Hamilton and Trolier, 1986; Bar-Tal et al., 2013). NLP models are prone to absorbing and amplifying these stereotypes (Bolukbasi et al., 2016; Zhao et al., 2017; Nangia et al., 2020), which can lead to algorithmic unfairness in downstream tasks like clinical note processing (Zhang et al., 2020). Contact notes frequently include direct quotes or paraphrases of statements made by others, and thus, they are liable to reflect not only possible implicit biases of the authors but also possible biases of anyone quoted.

This risk is particularly salient given decades of research have demonstrated racial disparities in the child welfare system (Roberts, 2009; Dettlaff et al., 2011; Hill, 2004; Wells et al., 2009; Roberts, 2019; of Health and Services, 2017). When faced with similarly severe child injuries, doctors may be more likely to report black families for abuse than white families (Lane et al., 2002; Roberts, 2009). One study found that black children were more likely to be placed in foster care even when controlling for risk factors such as parental substance abuse and allegations of abuse (Hill, 2005). Roberts (2019) asserts that black families of limited means are not only disproportionately involved in the system but are held to different standards of supervision. Institutional racism in child welfare has also been attributed to the links between the child welfare system and other systems like mental health services, criminal justice, and education (Hill, 2004). We investigate how models trained on contact notes may amplify these disparities.

Research Question 4 *Do contact notes encode racial biases that exacerbate algorithmic unfairness?*

Evaluation We characterize data imbalances and predictive disparities for black and white children, following the approach of Chouldechova et al. (2018). Specifically, we first examine if there are possible data biases by computing text and word statistics. We then investigate if possible data biases affect model predictions by comparing model performance and calibration for children of different races (Chouldechova et al., 2018).

5.2.2 Methodology

In this section, we develop and evaluate models that incorporate text and structured features. We briefly analyze and discuss the performance of different models, and we use the best-performing models in our subsequent analyses. We use two primary modeling variants, with different feature combinations:

Statistical Classifiers We first incorporate text features into models trained with structured data in prior work (Chouldechova et al., 2018). In these models, we use text-only features, structured-only features, and combined (hybrid) features:

- **Text Only**: We extract TF-IDF-weighted bag-of-words features using a 10,000 word vocabulary.
- **Structured Only**: We use the same 818 structured features as the AFST model

- **Hybrid**: Early experiments showed that concatenating the full 10,000 text features with the 818 structured features caused the model to ignore the structured features. Instead, we first train a text-only model using a logistic regression classifier. We then take the 500 words with the highest learned coefficients and the 500 words with the lowest (most negative) coefficients and construct TF-IDF features from this 1,000 word vocabulary. We then concatenate these features with the 818 structured features, constructing 1,818-dimensional feature vectors.

We show results using a random forest classifier with 500 trees and a logistic regression classifier, which is most similar to the AFST model currently in use.

NLP Models Second, we incorporate structured features into models designed for text data. We use two different NLP models developed for text-processing:

- **GatedCNN**: We use a state-of-the-art model designed for a similar task, assigning codes to medical notes (Ji et al., 2021). This model uses a CNN-based architecture that involves injecting word embedding between layers and an LSTM-style gating mechanism. The original model additionally computes dot-product interactions between medical notes and codes, using word embeddings derived from code descriptions. In our hybrid model, we incorporate structured features by replacing the medical code representations with the AFST structured features.

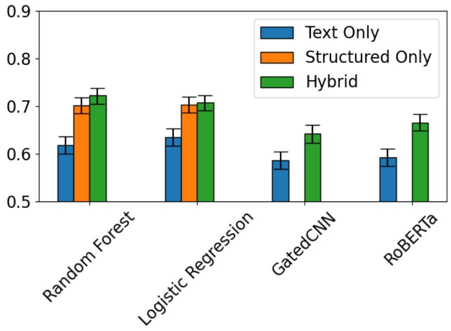
- **RoBERTa**: We train a RoBERTa-based classifier (Liu et al., 2019), where classification decisions are made using the final CLS representaiton. To incorporate structured features, we concatenate them to the CLS representation and pass the concatenated vector through a fully-connected linear layer, followed by a soft-max layer. Prior to training the model, we conduct additional masked-language-model pretraining over the full data set of 3.1M contact notes for 1 epoch, which prior work has shown improves performance on domain-specific data (Gururangan et al., 2020).

We do not have structured-only variants of the NLP models, as these models are designed for text features.

Experimental Set-up Our basic data unit is a referral-child pair: if a referral has multiple children, or if the same child is included on multiple referrals, we treat each unit as a separate data point. We base our models on the same data used to investigate the AFST model in prior work (Chouldechova et al., 2018). We keep the same test set, which was constructed to ensure there is no overlap in children or referrals between the train and test sets. We reserve 10% of the training set as a development set.

To identify relevant text data for each child on each referral, we pull contact notes for prior cases and referrals where the child is listed as a client, restricted to notes written within the previous 365 days that contain the first name of the child. We pre-process the text by expanding acronyms commonly used in contact notes using a manually curated list (e.g., "F → Father"). We further identify named entities using SpaCy[10] and we mask names of people and locations. We also remove first and last names of clients listed as active on cases and referrals at the time notes were written. While all data points contain structured features, not all families have prior interactions with the

[10] https://spacy.io/



Common Words		Highly Weighted Words	
Caseworker	Cas, ework, er	twin	tw, in
asked	ask, ed	lice	l, ice
Grandmother	Grand, mother	decision	dec, ision
Maternal	M, aternal	concern	con, cern
visit	vis, it	custody	c, ust, ody
concerns	con, cern, s	probation	pro, b, ation
Youth	Y, outh	unexcused	un, exc, used
Intake	Int, ake	paramour	param, our
Families	F, am, ilies	spank	sp, ank
Paternal	P, aternal	furniture	f, urn, iture
denied	den, ied		

Figure 5.6: Left: AUC scores of model variants. Metrics only include test data with text. 90% confidence intervals are computed using bootstrap sampling over the test set. Right: Words that RoBERTa tokenizer splits into to word pieces, includes the most common words in the training corpus that are split and the words assigned the highest or lowest weights by text-only logistic regression classifiers.

Department of Human Services, and thus not all data points contain associated text notes. Our final data set consists of 28,769 training instances, 7,893 of which contain text data, and 14,417 test data points, 4,133 of which contain text data. All text-only models are only trained on training data points that contain text. Structured and hybrid models are trained on all data points. We provide model hyperparameter settings in §A.2.2.

Model Performance Figure 5.6 reports AUC scores for the different models. The text-only models perform worse than the structured-only or hybrid models, and the GatedCNN and RoBERTa models perform worse than the bag-of-words statistical classifiers. Although neural models and pre-trained language models have achieved state-of-the-art performance on numerous NLP tasks, we observe several properties of our data that likely makes these models less useful in this setting. First, these parametric models require large data sets to achieve high-performing results, and it is likely that our data set is too small to train these models for a complex task. Second, our task is document-level, rather than sentence or paragraph level, where neural models typically excel. In order to make training tractable, it is necessary to truncate input text, which could drop useful information, even when truncation is done using heuristics (Hsu et al., 2020). Popular variants of the high-performing transformer architecture (including RoBERTa) in particular only support inputs up to 512 tokens (Liu et al., 2019). While some models use hierarchies of transformers or other connections to evade this, we found these models performed strictly worse on our data set—likely they require a larger training corpus (Pappagari et al., 2019). Finally, one of the primary advantage of neural models over bag-of-words approaches is their ability to capture syntax and interactions between words. However, contact notes have highly inconsistent formatting. While note-writers often use similar vocabulary, some write stream-of-consciousness style records of interactions, some follow a fixed structure with section headings, and some notes even consist of other documents like emails directly copy-and-pasted into the note. These differences of format often reflect the preferences of different supervisors and likely hinder the performance of models that capture syntactic information.

We additionally highlight the difficulty of using models pre-trained on other corpora in this setting. These models derive their vocabulary from the pre-training corpora, which can differ greatly from domain-specific settings. RoBERTa uses Byte-Pair Encoding (BPE) to enable handling large

Test set	Model	AUC	Avg. Pos Score	Avg. Neg Score	FPR	FNR
Full	Struct.	75.75 ± 0.02	13.85 ± 0.00	8.70 ± 0.00	19.58 ± 0.01	6.32 ± 0.01
	Hybrid	76.25 ± 0.02	13.94 ± 0.00	8.69 ± 0.00	19.52 ± 0.01	6.28 ± 0.01
Examples with notes	Struct.	69.79 ± 0.03	14.54 ± 0.01	11.21 ± 0.01	31.84 ± 0.05	7.92 ± 0.02
	Hybrid	71.83 ± 0.04	15.84 ± 0.01	13.25 ± 0.01	40.38 ± 0.09	5.87 ± 0.02

Table 5.3: Predictive performance of the structured and hybrid models in the supervised risk prediction task, where models are trained and tested on the same outcome, placement out-of-home. The feature inputs to the structured model is the tabular structured data, and the feature inputs to the hybrid is both the structured data and contact notes. Avg. Pos/Neg Score report the average predicted risk scores for true positive (placement occurred) and true negative (no placement) test data, where risk scores are computed by bucketing test predictions into ventiles. *Top:* Differences in model performance across the full test set ($n = 14,417$) are small. *Bottom:* Differences across the test set that contains text data ($n = 4,133$) show reductions in false negatives, but not in false positives.

vocabulary, which divides out-of-vocabulary words into sub-pieces in order to derive components that are part of the vocabulary (Liu et al., 2019; Radford et al., 2019; Sennrich et al., 2016). However, domain-specific vocabulary is often most informative. In Figure 5.6, we show that many of the words common in our training corpus and assigned high weights by bag-of-words classifiers are not included in RoBERTa's vocabulary. While some words are sub-divided into semantically coherent pieces like "grandmother" → "grand" "mother", others are less coherent "caseworker" → "cas" "ework", "er". Custom subword division could possible improve handling of out-of-vocabulary words (e.g. "caseworker" → "case" "work", "er") or methodology for incorporating new vocabulary words into pre-trained models. We note that additional masked-language-model training does not inherently change the model vocabulary, though it may improve the representations of common subwords. Based on the model performance results in Figure 5.6, in the remainder of this work we analyze the performance of the random forest structured and hybrid models.

5.2.3 Analysis of Opportunities and Pitfalls

In this section, we empirically investigate the risks and opportunities introduced in §5.2.1 by comparing the random forest hybrid and structured models. In order to both provide realistic estimates of how these systems may operate when deployed in practice and highlight differences in performance when incorporating text features, we report metrics over the full test set and over only the subset of the test data that contains text features. When computing metrics requiring a classification decision (e.g., FPR, FNR), we consider the 25% of test data with the highest raw output scores as having positive predictions. This percentage is consistent with prior work and corresponds to the mandatory screen-in threshold used by the Department of Human Services (Chouldechova et al., 2018). Additionally, given the small difference in overall AUC between the structured and hybrid models in Figure 5.6, we conduct bootstrap sampling over the training data, and report average metrics and standard error values computed from 100 models trained on different training data sample instances.

Potential opportunities Table 5.3 reports the overall performance results of the structured and hybrid models. The overall change in AUC and in predicting risks scores are small, though the hybrid model does consistently outperform the structured model across training samples. The improvements are largely driven by increases in the risk scores for data points where out-of-home placement occurred, which decrease the false negative rate. However, the incorporation of text features does not decrease the false positive rate for data points with text features, nor result in lower risk scores for data points where out-of-home placements did not occur. In general, the model interprets the presence of associated text as an indicator of risk, as test data points with text are assigned higher scores by the hybrid model than the structured model, regardless of whether or not out-of-home placement occurred. Overall, these results suggest that the incorporation of text data supports an affirmative answer to Question 1, but not to Question 2.

However, the results in Table 5.3 are imperfect metrics, as aggregate metrics can mask how model changes might affect individuals, and out-of-home placement is a proxy task reflective of a future caseworker's decision and not an objective measure of risk. In child welfare settings, a different prediction for even a single child could have an enormous impact. In order to investigate possible individual effects, we identify data points where there was the biggest difference in prediction score between the structured and hybrid models and manually examine associated contact notes.

In cases where the hybrid model predicted a lower score than the structured model, notes often contain descriptions of home visits and interviews with families where the writer observed no cause for concern. In the text-only logistic regression classifier (the best-performing text-only model, Figure 5.6), "concern" is one of the words the model assigns a low weight to, indicating it is predictive of no future placement.[11] In one case where the hybrid model predicts a lower score than the structured model, associated contact notes reveal that one person has frequently called the police and social services on a particular family as a way of targeting them, though prior investigations have revealed no causes for concern. This case offers an example of how free-form text notes can provide context that mitigate risk factors—while structured data would reveal that the family has been referred several times previously, which is typically predictive of future risk, the contact notes reveal prior allegations were unfounded and disingenuous. In cases where the hybrid model predicted a higher score than the structured model, notes often contain descriptions of concerns or unmet needs of families, such as needing help transporting children to school.

While aggregate differences in Table 5.3 are small, qualitative analysis suggests that there are individual cases where prior notes can aid in identifying risks and protective factors not captured in structured data. As call screen workers currently rarely examine prior notes when making screening decisions, one practical use case of the hybrid model could be determining when and which prior notes to consult manually during a call screening decision.

Pitfalls We investigate Question 3, whether text data is liable to increase overfitting to human decisions and proxy tasks, by examining how predictive contact notes are of near-term decisions and how models trained to predict out-of-home placements transfer to other adverse events. Figure 5.7 reports the AUC scores of the random forest structured-only and hybrid models when contact notes associated with referrals are used to predict whether or not the referral will be accepted for services and out-of-home placements. As we use a different data set of contact notes for this task (notes associated with current referrals, frequently generated during investigations), we report details on the data in §A.2.1. The largest performance improvement occurs when predicting the service decision

[11] While this finding may seem counter-intuitive at first glance, when there is cause for concern, note-writers generally uses more detailed language to describes the issue (e.g., substance abuse). The generic "concern" is mostly used in the context of "no cause for concern."

Model	Service AUC	Place. AUC
Struct.	77.2	75.9
Hybrid	86.1	80.5

Highest-weighted Words	
Service	crisi, servic, hear, decis, group
Placement	foster, placement, hear, author, physic

Figure 5.7: *Left:* AUC for random forest models with and without text features from notes directly associated with referrals over two nearer-term decisions: whether or not the referral will be accepted for services and out-of-home placement. Metrics only include test data points that have text features (Table A.5 reports data sizes). *Right:* words assigned the highest weights by a text-only logistic regression classifier. Notes often document and explicitly justify decisions.

			Avg.	Avg.		
Test Set	Model	AUC	Pos Score	Neg Score	FPR	FNR
Full	Struct.	68.94 ± 0.02	12.77 ± 0.00	8.98 ± 0.00	22.12 ± 0.01	6.80 ± 0.01
	Hybrid	69.26 ± 0.03	12.83 ± 0.00	8.97 ± 0.00	21.74 ± 0.01	6.49 ± 0.01
Examples with notes	Struct.	66.07 ± 0.03	14.16 ± 0.01	11.33 ± 0.01	33.23 ± 0.05	8.54 ± 0.02
	Hybrid	66.99 ± 0.05	15.42 ± 0.01	13.38 ± 0.01	41.50 ± 0.08	6.33 ± 0.03

Table 5.4: Predictive performance of the structured and hybrid models in the task transfer risk prediction task, where the models are trained on placement and tested on an aggregated indicator of other adverse outcomes. Top includes all test data (14,417). Bottom includes only test data with text features (4,133). We report details on the construction of the aggregated indicator in Table A.6.

of a referral using notes on the current referral. A text-only logistic regression classifier over this same task assigns the highest weights to *decis* (lemmatized *decision*), *hear*, *servic* (lemmatized *service*), *crisi* (lemmatized *crisis*), demonstrating that these notes often explicitly document the intended service decision. Similarly, text features improve AUC when predicting future placement using notes from a current referral. The top-weighted words in a text-only model for this task are *placement* and *foster*, suggesting that notes often document intention or initial steps for out-of-home placements.

Overall, these results support the hypothesis that notes largely reflect writers' impression of events and often contain explicit documentation of decisions made on cases and referrals. While notes can provide new information, reliance on them in decision-making risks reinforcing existing viewpoints and decisions. An ideal text processing system would separate text content that documents and justifies decisions from content focusing on raw observations, though this is likely impossible in practice, as even when not explicitly describing decisions, note-writers are likely to document observations that support their existing opinions.

In Table 5.4, we report results of the task transfer risk prediction task: how well a model trained to predict out-of-home placement is able to predict an aggregated indicator of other adverse outcomes, including behavioral or mental health concerns, housing support, or involvement in court cases. The hybrid model performs no worse than the structured model, and performs better on some metrics. Most importantly, the hybrid model does not predict lower risk scores for true positives than the structured model. We do not therefore find evidence that it underestimates risk of adverse outcomes it was not explicitly trained to estimate when compared with the structured model. Thus, while we do find that text data is liable to reinforcing existing viewpoints and decisions, we do not find evidence that it increases over-fitting to the proxy task.

Finally, we investigate Question 4, whether text features increase algorithmic unfairness and exacerbate racial biases. We first consider data size and word statistics, in order to investigate

		Avg.	Avg.
	Num.	| Prev |	| Curr |
Black	20,527	634.8	2,497.2
White	17,726	840.5	2,360.8
Unknown	3,540	363.1	2,018.6
Other	3,331	787.9	2,446.8
Mixed	1,382	912.2	2,593.3

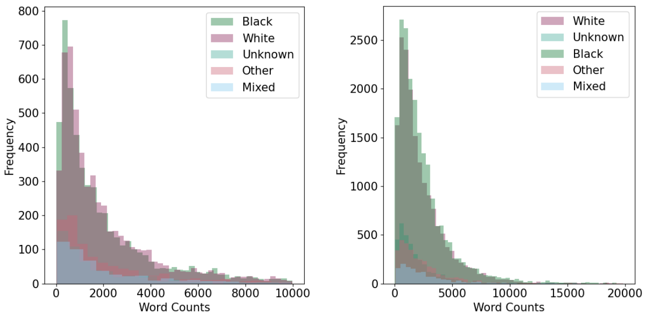


Figure 5.8: Length and word statistics of notes associated with cases and referrals for the AFST task by race. Left: the number of data points, average word count in *previous notes* preceding the current referral (our primary data set), and average word count in *current notes* directly associated with the current referral. Center: histogram of word counts in *previous notes* (excluding data points with 0 notes for readability). Right: histogram of word counts in *current notes*. There are not consistent differences in statistics for notes associated with black children and white children.

if there are reasons to believe text data differs for children of different races. Figure 5.8 reports the average number of words and word-count histograms in contact notes associated with children of different races on referrals. We report metrics for both notes occurring before each referral (e.g. *previous notes*, which are incorporated into our primary hybrid model), and for *current notes* directly associated with each referral, typically generated during the investigation phase. There are not consistent differences in the amount of data available for children of each race (e.g., on average, there is more text data in *previous notes* for white children, but more in *current notes* for black children).

While we do not find consistent length differences, in Table 5.5, we use word statistics to examine possible content differences in notes about black and white children. More specifically, for all words in the data, we compute to what extent each word is over-represented in notes about black children or white children as compared to all other notes using log-odds with a Dirichlet prior (Monroe et al., 2008). In order to compare these over-represented words with common alternatives, we train 100-dimensional word embeddings from the full data set of 3.1M contact notes using skip-gram Word2Vec with a context window of 5. For each of the 100 most-overrepresented words, we identify the 3 words with the most similar word embeddings that are not overrepresented (e.g. log-odds score < 0), discarding words that occur < 20 times in the data set. Table 5.5 displays a subset of these 100 words and their 3 nearest neighbors for black and white associations. These word associations reveal possible content and style differences along racial lines. Words common in notes about black children focus on behavior and punishment, while words common in notes about white children focus on drug use. Additionally, some near-synonyms have racial associations: notes about black children use “whooped” over “spanked”, “informed” over “reassured”, and “disrespectful” over “rude”. These differences likely reflect both terminology used by note writers as well as terminology used by the clients and sources they interview. Differences in language variants can lead to to harmful biases in downstream tasks, e.g., off-the-shelf NLP models for toxic language classification are more likely to falsely classify African American English as offensive (Davidson et al., 2019; Sap et al., 2019; Field et al., 2021). While more research is needed to investigate the origin and effects of these differences in content and terminology, these word statistics suggest that there are systemic differences in notes

Black-assoc.	Score	Nearest Match	White-assoc.	Score	Nearest match
she	59.1	he,m,mgm	father	34.0	mother,fathers,grandmother
her	52.9	his,hers,them	heroin	33.8	marijuana,ecstasy,thc
belt	35.7	spatula,paddle,spoon	he	31.2	she,c
suspended	25.2	absent,bullied,grounded	drugs	25.6	marijuana,weed,k2
informed	23.4	reassured,clarified,infomred	anxiety	24.9	paranoia,postpartum,anorexia
whooped	22.9	spanked,paddled,smacked	using	21.3	smoking,utilizing,craving
punishment	19.0	grounded,electronics,consequence	sick	20.5	congested,tired,hungry
placed	18.9	residing,removed,released	grandparents	19.0	grandmother,aunt,aunts
disrespectful	16.7	rude,defiant,mouthy	rehab	19.0	ars,prison,ridgeview
curfew	15.7	bedtime,awoled,consequence	reports	17.7	stated,indicated,reprots
beds	15.7	cribs,bunk,blankets	grounded	16.3	punished,punishment,disciplined

Table 5.5: Words overrepresented in notes relevant for black children and white children computed using log-odds with a Dirichlet prior (Monroe et al., 2008). For reference, nearest-matching words that are not associated with the specified race are identified using cosine similarity of word embeddings. There are noticeable differences in topics and terminology in notes about children of different races.

			Avg. Pos	Avg. Neg		
Test Set	Model	AUC$_{black-white}$	Score$_{b-w}$	Score$_{b-w}$	FPR$_{b-w}$	FNR$_{b-w}$
Full	Struct.	-0.00 ± 0.00	0.35 ± 0.01	0.64 ± 0.01	0.01 ± 0.00	0.02 ± 0.00
	Hybrid	0.01 ± 0.00	0.33 ± 0.01	0.30 ± 0.01	0.00 ± 0.00	0.02 ± 0.00
w/ notes	Struct.	-0.02 ± 0.00	0.14 ± 0.01	0.71 ± 0.01	0.03 ± 0.00	0.03 ± 0.00
	Hybrid	-0.01 ± 0.00	0.32 ± 0.01	0.67 ± 0.01	0.06 ± 0.00	0.02 ± 0.00

Table 5.6: Disparities in predictive performance metrics by race for the supervised and hybrid models. The predictive disparities are largely comparable for the structured and hybrid models. We do not therefore find evidence that incorporation of text features increases aggregate measures of algorithmic unfairness in this setting. Raw performance values are reported in §A.2.3.

about children of different races which could be absorbed and amplified by NLP models.

We next directly measure if these observed differences in text increase algorithmic unfairness in the hybrid model by comparing accuracy equity and error rates and calibration (Dieterich et al., 2016; Chouldechova et al., 2018). Table 5.6 reports differences in structured model performance for black and white children. Differences are small overall, and the hybrid model does not show greater performance disparities than the structured model. In identifying referrals with future out-of-home placement (e.g. FNR), the hybrid model slight reduces disparities. In identifying referrals without out-of-home placements (e.g. FPR), the hybrid model slightly improves performance for black children more than white children.

Figure 5.9 compares risk scores using calibration plots. Both the structured and hybrid models display signs of miscalibration in the highest risk bracket, which is consistent with findings in prior work (Chouldechova et al., 2018). More specifically, a higher percentage of black children assigned the highest risk eventually are placed out-of-home than white children, but the hybrid model does not appear any more miscalibrated than the structured model.

While the results in Table 5.6 and Figure 5.9 do not show signs of increased racial disparities, both of these figures compute results based on out-of-home placement values. As it is possible that this proxy outcome itself reflects racial bias (See Section 5.2.1), we additionally consider disparate impact-type metrics that do not depend on this proxy outcome. Figure 5.10 displays the percent of children of each race that are flagged as high-risk as well as the racial composition of those that are flagged under the hybrid and structured models. These metrics allow us to examine if the hybrid

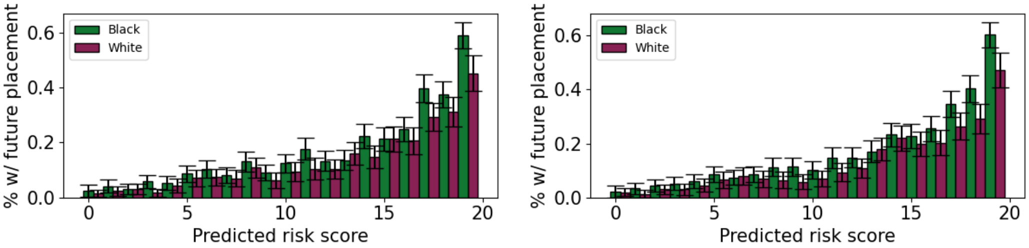


Figure 5.9: Calibration plots for AFST model using structured (left) and hybrid (right) features. We infer predicted risk scores by grouping the full data set into ventiles, but only display data points for black and white children in these figures. Both the structured and hybrid models display signs of miscalibration in the highest ventile. We refer to Chouldechova et al. (2018) for details on computing and interpreting calibration plots.

Model	$\%race_{b-w}$	$\%flagged_{b-w}$
Struct.	3.27 ± 0.06	13.46 ± 0.11
Hybrid	2.76 ± 0.06	12.57 ± 0.11

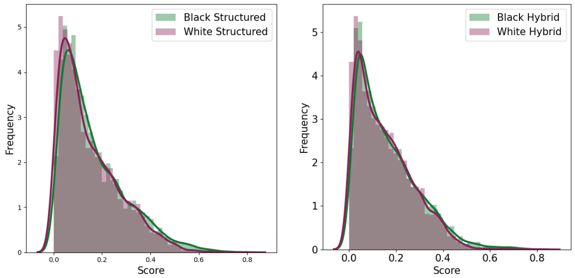


Figure 5.10: *Left:* difference in the percent of children of each race (black - white) that are flagged as high-risk under the specified model (%race), difference in the racial composition (black - white) of those who are flagged as high risk (%flagged). The hybrid model shows a reduced disparity in the percentage of black children who are flagged as high-risk. *Right:* histograms of the raw scores outputted by each model, divided by race. Under the structured model, score predictions for black children are right-shifted compared to white children. Under the hybrid model, the score distributions are nearly identical.

model is liable to increase the number of black families involved in child welfare services.

While under both models, a higher percentage of black children are flagged as high-risk relative to white children, the hybrid model reduces this disparity. This reduction is also visible in the histograms of the raw scores outputted by the model: under the structured model, score predictions for black children are right-shifted compared to white children, while under the hybrid model, the score distributions are nearly identical (Figure 5.10, right).

The changes in model performance when text features are incorporated suggest that text features do not increase aggregate measures of racial disparities in model predictions and may actually improve them when compared to a structured-only model. While text data is liable to perpetuate bias and stereotypes, it can also provide context important for disambiguating risk factors in structured data that have differential validity by race. For instance due to systemic racism in the U.S. criminal justice system, a criminal record is a much more salient indicator of risk for a white parent than a black parent (Alexander, 2011). Nevertheless, we emphasize that this finding should not be interpreted as signaling a lack of racial bias in the contact notes or in the predictive models used on those notes. It may be that a combination of divergent factors are at play. Our analysis of racial disparities points to the complicated conclusion that while text data may introduce new biases, it might also partially address flaws in the structured data.

Overall, we find little evidence that incorporating text features into an existing child welfare risk assessment offers widespread benefits, but we do find evidence that the targeted processing and use of certain contact notes may identify pertinent risk signals and protective factors that could improve decision-making in certain cases. Our research highlights how it is difficult to disambiguate content that reflects individual viewpoints and decisions from content that introduces valuable information. These findings suggest that, rather than their use in automated predictive risk models, contact notes may be better suited for use in a more transparent system. Future research should consider how predictive models might be useful for identifying which notes should be read manually or how tools for information extraction or targeted summarization could help distill the most informative content. Progress in this area will require significant advances in processing domain-specific terminology and preventing racialized differences in notes from manifesting in biased and unfair models. We recommend that future work include qualitative methods to evaluate performance and fairness as our analysis underscores the limitations of aggregate quantitative metrics. Finally, we call for future research to engage with stakeholders including affected families, caseworkers, and community advocates to understand and integrate their perspectives on the use of the fully or semi-automated processing of case notes in decision-making.

5.3 Conclusions and Future Work

There has been much interest in the potential for using machine learning models to inform policy decisions. This chapter presents two case studies of how NLP analyses can inform organizers of social movements and provide insight into the risks and benefits of relying on text-processing systems. These studies expose some of the limitations of current NLP systems, which make them difficult to adapt to real-world settings. Both reveal that not only is text data is often domain-specific, but also that the most important content is often the most domain-specific. In §5.1.2 we note how off-the-shelf lexicons associate “police” with trust, which is contradictory to the concept of police in protests against police brutality. Similarly, in §5.2.2, we show that many critical vocabulary words in child welfare settings are not in RoBERTa’s vocabulary. These examples call for future research in domain-adaptation approaches, including ways of updating off-the-shelf model vocabulary and encouraging models to unlearn associations from training data that may not hold in application settings.

Within each project, there are also numerous directions for future work. There is a long history of sentiment analysis in NLP, with a substantial amount of work also investigating expressions of emotions. However, the emotions relevant for specific scenarios can vary. While we use Ekman’s taxonomy in order to leverage existing annotations (Demszky et al., 2020), a finer-grained breakdown of positive emotions would be valuable, e.g. distinguishing optimism, solidarity, and pride. Additionally, we focus on a single month of high volatility, but a longer-term study could likely lead to more definitive conclusions about the timing of social media posts and offline protests and whether or not specific emotions expressed on social media precede protests with high turnout.

§5.2.2 similarly only addresses one specific use case of language technologies in child welfare settings. There are numerous other avenues where NLP could aid childwelfare agencies in serving their clients. For example, cases often span years and can have thousands of notes. Tools for extracting information and tracking incidents over time can help caseworkers keep track of events and quickly find relevant information, especially in the event of an emergency. Similarly, summarization tools can help someone reviewing a case for the first time, such as a new caseworker or a supervisor, quickly get up to speed. NLP systems to identify signs of implicit bias could also be useful for mitigating observed racial disparities in the child welfare system and providing guidance on how the system might be changed. However, all of the technology discussed in this chapter has the potential to be misused and could result in increased stereotyping or amplification of historical bias. §7 provides a more in-depth discussion of the ethical implications of this work.

Chapter 6

Case Examination of the Risks and Harms of NLP

The chapter discusses work previously published in (Field et al., 2021).

While the preceding chapters primarily discuss the potential of NLP in characterizing and mitigating social issues, this chapter takes a more critical perspective and highlights ways that NLP can perpetuate, amplify, and create harms. Although there are numerous ways that well-intentioned research in machine learning and AI can have harmful consequences, this chapter presents a case examination through one particular dimension: racial injustice.

Race and language are tied in complicated ways. Raciolinguistics scholars have studied how they are mutually constructed: historically, colonial powers construct linguistic and racial hierarchies to justify violence, and currently, beliefs about the inferiority of racialized people's language practices continue to justify social and economic exclusion (Rosa and Flores, 2017).[1] Furthermore, language is the primary means through which stereotypes and prejudices are communicated and perpetuated (Hamilton and Trolier, 1986; Bar-Tal et al., 2013).

However, questions of race and racial bias have been minimally explored in NLP literature. While researchers and activists have increasingly drawn attention to racism in computer science and academia, frequently-cited examples of racial bias in AI are often drawn from disciplines other than NLP, such as computer vision (facial recognition) (Buolamwini and Gebru, 2018) or machine learning (recidivism risk prediction) (Angwin et al., 2019). Even the presence of racial biases in search engines like Google (Sweeney, 2013; Noble, 2018) has prompted little investigation in the ACL community. Work on NLP and race remains sparse, particularly in contrast to concerns about gender bias, which have led to surveys, workshops, and shared tasks (Sun et al., 2019; Webster et al., 2019).

This chapter presents a comprehensive survey of how NLP literature and research practices engage with race. We first examine 79 papers from the ACL Anthology that mention the words 'race', 'racial', or 'racism' and highlight examples of how racial biases manifest at all stages of NLP model pipelines (§6.2). We then describe some of the limitations of current work (§6.3), specifically

[1] We use *racialization* to refer the process of "ascribing and prescribing a racial category or classification to an individual or group of people . . . based on racial attributes including but not limited to cultural and social history, physical features, and skin color" (Hudley, 2017).

showing that NLP research has only examined race in a narrow range of tasks with limited or no social context. Finally, in §6.4, we revisit the NLP pipeline with a focus on how *people* generate data, build models, and are affected by deployed systems, and we highlight current failures to engage with people traditionally underrepresented in STEM and academia.

While little work has examined the role of race in NLP specifically, prior work has discussed race in related fields, including human-computer interaction (HCI) (Ogbonnaya-Ogburu et al., 2020; Rankin and Thomas, 2019; Schlesinger et al., 2017), fairness in machine learning (Hanna et al., 2020), and linguistics (Hudley et al., 2020; Motha, 2020). We draw comparisons and guidance from this work and show its relevance to NLP research. Our work differs from NLP-focused related work on gender bias (Sun et al., 2019), 'bias' generally (Blodgett et al., 2020), and the adverse impacts of language models (Bender et al., 2021) in its explicit focus on race and racism.

In surveying research in NLP and related fields, we ultimately find that *NLP systems and research practices produce differences along racialized lines.* Our work calls for NLP researchers to consider the social hierarchies upheld and exacerbated by NLP research and to shift the field toward "greater inclusion and racial justice" (Hudley et al., 2020).

6.1 Background

It has been widely accepted by social scientists that race is a social construct, meaning it "was brought into existence or shaped by historical events, social forces, political power, and/or colonial conquest" rather than reflecting biological or 'natural' differences (Hanna et al., 2020). More recent work has criticized the "social construction" theory as circular and rooted in academic discourse, and instead referred to race as "colonial constituted practices", including "an inherited western, modern-colonial practice of violence, assemblage, superordination, exploitation and segregation" (Saucier et al., 2016).

The term race is also multi-dimensional and can refer to a variety of different perspectives, including *racial identity* (how you self-identify), *observed race* (the race others perceive you to be), and *reflected race* (the race you believe others perceive you to be) (Roth, 2016; Hanna et al., 2020; Ogbonnaya-Ogburu et al., 2020). Racial categorizations often differ across dimensions and depend on the defined categorization schema. For example, the United States census considers Hispanic an ethnicity, not a race, but surveys suggest that 2/3 of people who identify as Hispanic consider it a part of their racial background.[2] Similarly, the census does not consider 'Jewish' a race, but some NLP work considers anti-Semitism a form of racism (Hasanuzzaman et al., 2017). Race depends on historical and social context—there are no 'ground truth' labels or categories (Roth, 2016).

As the work we survey primarily focuses on the United States, our analysis similarly focuses on the U.S. However, as race and racism are global constructs, some aspects of our analysis are applicable to other contexts. We suggest that future studies on racialization in NLP ground their analysis in the appropriate geo-cultural context, which may result in findings or analyses that differ from our work.

We also acknowledge that we, the authors of this work, are situated in the cultural contexts of the United States of America and the United Kingdom/Europe. We all identify as NLP researchers,

[2] https://www.census.gov/mso/www/training/pdf/race-ethnicity-onepager.pdf/, https://www.census.gov/topics/population/race/about.html, https://www.pewresearch.org/fact-tank/2015/06/15/is-being-hispanic-a-matter-of-race-ethnicity-or-both/

and we acknowledge that we are situated within the traditionally exclusionary practices of academic research. These perspectives have impacted our work, and there are viewpoints outside of our institutions and experiences that our work may not fully represent.

	Collect Corpus	Analyze Corpus	Develop Model	Detect Bias	Debias	Survey/Position	Total
Abusive Language	6	4	2	5	2	2	21
Social Science/Social Media	2	10	6	1	-	1	20
Text Representations (LMs, embeddings)	-	2	-	9	2	-	13
Text Generation (dialogue, image captions, story gen.)	-	-	1	5	1	1	8
Sector-specific NLP applications (edu., law, health)	1	2	-	-	1	3	7
Ethics/Task-independent Bias	1	-	1	1	1	2	6
Core NLP Applications (parsing, NLI, IE)	1	-	1	1	1	-	4
Total	11	18	11	22	8	9	79

Table 6.1: 79 papers on race or racism from the ACL anthology, categorized by NLP application and focal task.

6.2 Survey of NLP literature on race

6.2.1 ACL Anthology papers about race

In this section, we introduce our primary survey data—papers from the ACL Anthology[3]—and we describe some of their major findings to emphasize that *NLP systems encode racial biases.* We searched the anthology for papers containing the terms 'racial', 'racism', or 'race', discarding ones that only mentioned race in the references section or in data examples and adding related papers cited by the initial set if they were also in the ACL Anthology. In using keyword searches, we focus on papers that explicitly mention race and consider papers that use euphemistic terms to not have substantial engagement on this topic. As our focus is on NLP and the ACL community, we do not include NLP-related papers published in other venues in the reported metrics (e.g. Table 6.1), but we do draw from them throughout our analysis.

Our initial search identified 165 papers. However, reviewing all of them revealed that many do not deeply engage on the topic. For example, 37 papers mention 'racism' as a form of abusive language or use 'racist' as an offensive/hate speech label without further engagement. 30 papers only mention race as future work, related work, or motivation, e.g. in a survey about gender bias, "Non-binary genders as well as racial biases have largely been ignored in NLP" (Sun et al., 2019). After discarding these types of papers, our final analysis set consists of 79 papers.[4]

Table 6.1 provides an overview of the 79 papers, manually coded for each paper's primary NLP task and its focal goal or contribution. We determined task/application labels through an iterative

[3] The ACL Anthology includes papers from all official ACL venues and some non-ACL events. As of December 2020 it included 6, 200 papers

[4] We do not discard all papers about abusive language, only ones that exclusively use racism/racist as a classification label. We retain papers with further engagement, e.g. discussions of how to define racism or identification of racial bias in hate speech classifiers.

process: listing the main focus of each paper and then collapsing similar categories. In cases where papers could rightfully be included in multiple categories, we assign them to the best-matching one based on stated contributions and the percentage of the paper devoted to each possible category. We refer to Field et al. (2021) for additional categorizations of the papers according to publication year, venue, and racial categories used, as well as the full list of 79 papers.

6.2.2 NLP systems encode racial bias

Next, we present examples that identify racial bias in NLP models, focusing on 5 parts of a standard NLP pipeline: data, data labels, models, model outputs, and social analyses of outputs. We include papers described in Table 6.1 and also relevant literature beyond the ACL Anthology (e.g. NeurIPS, PNAS, Science). These examples are not intended to be exhaustive, and in §6.3 we describe some of the ways that NLP literature has failed to engage with race, but nevertheless, we present them as evidence that *NLP systems perpetuate harmful biases along racialized lines.*

Data A substantial amount of prior work has already shown how NLP systems, especially word embeddings and language models, can absorb and amplify social biases in data sets (Bolukbasi et al., 2016; Zhao et al., 2017). While most work focuses on gender bias, some work has made similar observations about racial bias (Rudinger et al., 2017; Garg et al., 2018; Kurita et al., 2019). These studies focus on *how* training data might describe racial minorities in biased ways, for example, by examining words associated with terms like ‘black’ or traditionally European/African American names (Caliskan et al., 2017; Manzini et al., 2019). Some studies additionally capture *who* is described, revealing under-representation in training data, sometimes tangentially to primary research questions: Rudinger et al. (2017) suggest that gender bias may be easier to identify than racial or ethnic bias in Natural Language Inference data sets because of data sparsity, and Caliskan et al. (2017) alter the Implicit Association Test stimuli that they use to measure biases in word embeddings because some African American names were not frequent enough in their corpora.

An equally important consideration, in addition to whom the data describes is *who authored the data*. For example, Blodgett et al. (2018) show that parsing systems trained on White Mainstream American English perform poorly on African American English (AAE).[5] In a more general example, Wikipedia has become a popular data source for many NLP tasks. However, surveys suggest that Wikipedia editors are primarily from white-majority countries,[6] and several initiatives have pointed out systemic racial biases in Wikipedia coverage (Adams et al., 2019; Field et al., 2022).[7] Models trained on these data only learn to process the type of text generated by these users, and further, only learn information about the topics these users are interested in. The *representativeness* of data sets is a well-discussed issue in social-oriented tasks, like inferring public opinion (Olteanu et al., 2019), but this issue is also an important consideration in ‘neutral’ tasks like parsing (Waseem et al., 2021). The type of data that researchers choose to train their models on does not just affect what *data* the models perform well for, it affects what *people* the models work for. NLP researchers cannot assume models will be useful or function for marginalized people unless they are trained on data generated by them.

[5] We note that conceptualizations of AAE and the accompanying terminology for the variety have shifted considerably in the last half century; see King (2020) for an overview.

[6] https://meta.wikimedia.org/wiki/Research:Wikipedia_Editors_Survey_2011_April

[7] https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia

Data Labels Although model biases are often blamed on raw data, several of the papers we survey identify biases in the way researchers categorize or obtain data annotations. For example:

- **Annotation schema** Returning to Blodgett et al. (2018), this work defines new parsing standards for AAE, demonstrating how parsing labels themselves were not designed for racialized language varieties.
- **Annotation instructions** Sap et al. (2019) show that annotators are less likely to label tweets using AAE as offensive if they are told the likely language varieties of the tweets. Thus, how annotation schemes are designed (e.g. what contextual information is provided) can impact annotators' decisions, and failing to provide sufficient context can result in racial biases.
- **Annotator selection** Waseem (2016) show that feminist/anti-racist activists assign different offensive language labels to tweets than figure-eight workers, demonstrating that annotators' lived experiences affect data annotations.

Models Some papers have found evidence that model instances or architectures can change the racial biases of outputs produced by the model. Sommerauer and Fokkens (2019) find that the word embedding associations around words like 'race' and 'racial' change not only depending on the model architecture used to train embeddings, but also on the specific model *instance* used to extract them, perhaps because of differing random seeds. Kiritchenko and Mohammad (2018) examine gender and race biases in 200 sentiment analysis systems submitted to a shared task and find different levels of bias in different systems. As the training data for the shared task was standardized, all models were trained on the same data. However, participants could have used external training data or pre-trained embeddings, so a more detailed investigation of results is needed to ascertain which factors most contribute to disparate performance.

Model Outputs Several papers focus on model outcomes, and how NLP systems could perpetuate and amplify bias if they are deployed:

- Classifiers trained on common abusive language data sets are more likely to label tweets containing characteristics of AAE as offensive (Davidson et al., 2019; Sap et al., 2019).
- Classifiers for abusive language are more likely to label text containing identity terms like 'black' as offensive (Dixon et al., 2018).
- GPT outputs text with more negative sentiment when prompted with AAE-like inputs (Groenwold et al., 2020).

Social Analyses of Outputs While the examples in this section primarily focus on racial biases in trained NLP systems, other work (e.g. included in 'Social Science/Social Media' in Table 6.1) uses NLP tools to analyze race in society. Examples include examining how commentators describe football players of different races (Merullo et al., 2019) or how words like 'prejudice' have changed meaning over time (Vylomova et al., 2019).

While differing in goals, this work is often susceptible to the same pitfalls as other NLP tasks. One area requiring particular caution is in the interpretation of results produced by analysis models. For example, while word embeddings have become a common way to measure semantic change or estimate word meanings (Garg et al., 2018), Joseph and Morgan (2020) show that embedding associations do not always correlate with human opinions; in particular, correlations are stronger for beliefs about gender than race. Relatedly, in HCI, the recognition that authors' own biases can affect their interpretations of results has caused some authors to provide self-disclosures (Schlesinger et al., 2017), but this practice is uncommon in NLP.

We conclude this section by observing that when researchers have looked for racial biases in NLP systems, they have usually found them. This literature calls for proactive approaches in considering how data is collected, annotated, used, and interpreted to prevent NLP systems from exacerbating historical racial hierarchies.

6.3 Limitations in where and how NLP operationalizes race

While §6.2 demonstrates ways that NLP systems encode racial biases, we next identify gaps and limitations in how these works have examined racism, focusing on *how* and *in what tasks* researchers have considered race. We ultimately conclude that prior NLP literature has marginalized research on race and encourage deeper engagement with other fields, critical views of simplified classification schema, and broader application scope in future work (Blodgett et al., 2020; Hanna et al., 2020).

6.3.1 Common data sets are narrow in scope

The papers we surveyed suggest that research on race in NLP has used a very limited range of data sets, which fails to account for the multi-dimensionality of race and simplifications inherent in classification. We identified 3 common data sources:

- 9 papers use a set of tweets with inferred probabilistic topic labels based on alignment with U.S. census race/ethnicity groups (or the provided inference model) (Blodgett et al., 2016).
- 11 papers use lists of names drawn from Sweeney (2013), Caliskan et al. (2017), or Garg et al. (2018). Most commonly, 6 papers use African/European American names from the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), which in turn draws data from Greenwald et al. (1998) and Bertrand and Mullainathan (2004).
- 10 papers use explicit keywords like 'Black woman', often placed in templates like "I am a ______" to test if model performance remains the same for different identity terms.

While these commonly-used data sets can identify performance disparities, they only capture a narrow subset of the multiple dimensions of race (§6.1). For example, none of them capture self-identified race. While observed race is often appropriate for examining discrimination and some types of disparities, it is impossible to assess potential harms and benefits of NLP systems without assessing their performance over text generated by and directed to people of different races. The corpus from Blodgett et al. (2016) does serve as a starting point and forms the basis of most current work assessing performance gaps in NLP models (Sap et al., 2019; Blodgett et al., 2018; Xia et al., 2020; Xu et al., 2019; Groenwold et al., 2020), but even this corpus is explicitly not intended to infer race.

Furthermore, names and hand-selected identity terms are not sufficient for uncovering model bias. De-Arteaga et al. (2019) show this in examining gender bias in occupation classification: when overt indicators like names and pronouns are scrubbed from the data, performance gaps and potential allocational harms still remain. Names also generalize poorly. While identity terms can be examined across languages (van Miltenburg et al., 2017), differences in naming conventions often do not translate, leading some studies to omit examining racial bias in non-English languages (Lauscher and Glavaš, 2019). Even within English, names often fail to generalize across domains, geographies, and time. For example, names drawn from the U.S. census generalize poorly to Twitter (Wood-Doughty et al., 2018), and names common among Black and white children were not distinctly different prior to the 1970s (Fryer Jr and Levitt, 2004; Sweeney, 2013).

We focus on these 3 data sets as they were most common in the papers we surveyed, but we note that others exist. Preoţiuc-Pietro and Ungar (2018) provide a data set of tweets with self-identified race of their authors, though it is little used in subsequent work and focused on demographic prediction, rather than evaluating model performance gaps. Two recently-released data sets (Nadeem et al., 2021; Nangia et al., 2020) provide crowd-sourced pairs of more- and less-stereotypical text. More work is needed to understand any privacy concerns and the strengths and limitations of these data (Blodgett et al., 2021). Additionally, some papers collect domain-specific data, such as self-reported race in an online community (Loveys et al., 2018), or crowd-sourced annotations of perceived race of football players (Merullo et al., 2019). While these works offer clear contextualization, it is difficult to use these data sets to address other research questions.

6.3.2 Classification schemes operationalize race as a fixed, single-dimensional U.S.-census label

Work that uses the same few data sets inevitably also uses the same few classification schemes, often without justification. The most common explicitly stated source of racial categories is the U.S. census, which reflects the general trend of U.S.-centrism in NLP research (the vast majority of work we surveyed also focused on English). While census categories are sometimes appropriate, repeated use of classification schemes and accompanying data sets without considering who defined these schemes and whether or not they are appropriate for the current context risks perpetuating the misconception that race is ‘natural’ across geo-cultural contexts. We refer to Hanna et al. (2020) for a more thorough overview of the harms of “widespread uncritical adoption of racial categories,” which “can in turn re-entrench systems of racial stratification which give rise to real health and social inequalities.” At best, the way race has been operationalized in NLP research is only capable of examining a narrow subset of potential harms. At worst, it risks reinforcing racism by presenting racial divisions as natural, rather than the product of social and historical context (Bowker and Star, 2000).

As an example of questioning who devised racial categories and for what purpose, we consider the pattern of re-using names from Greenwald et al. (1998), who describe their data as sets of names “judged by introductory psychology students to be more likely to belong to White Americans than to Black Americans” or vice versa. When incorporating this data into WEAT, Caliskan et al. (2017) discard some judged African American names as too infrequent in their embedding data. Work subsequently drawing from WEAT makes no mention of the discarded names nor contains much discussion of how the data was generated and whether or not names judged to be white or Black by introductory psychology students in 1998 are an appropriate benchmark for the studied task. While gathering data to examine race in NLP is challenging, and in this work we ourselves draw from examples that use Greenwald et al. (1998), it is difficult to interpret what implications arise when models exhibit disparities over this data and to what extent models without disparities can be considered ‘debiased’.

Finally, almost all of the work we examined conducts single-dimensional analyses, e.g. focus on race or gender but not both simultaneously. This focus contrasts with the concept of *intersectionality*, which has shown that examining discrimination along a single axis fails to capture the experiences of people who face marginalization along multiple axes. For example, consideration of race often emphasizes the experience of gender-privileged people (e.g. Black men), while consideration of gender emphasizes the experience of race-privileged people (e.g. white women). Neither reflect the experience of people who face discrimination along both axes (e.g. Black women) (Crenshaw, 1989). A small selection of papers have examined intersectional biases in embeddings or word co-occurrences (Herbelot et al., 2012; May et al., 2019; Tan and Celis, 2019; Lepori, 2020), but we did not identify mentions of intersectionality in any other NLP research areas. Further, several of these papers use NLP technology to examine or validate theories on intersectionality; they do not draw from theory on intersectionality to critically examine NLP models. These omissions can mask harms: Jiang and Fellbaum (2020) provide an example using word embeddings of how failing to consider intersectionality can render invisible people marginalized in multiple ways. Numerous directions remain for exploration, such as how 'debiasing' models along one social dimension affects other dimensions. Surveys in HCI offer further frameworks on how to incorporate identity and intersectionality into computational research (Schlesinger et al., 2017; Rankin and Thomas, 2019).

6.3.3 NLP research on race is restricted to specific tasks and applications

Finally, Table 6.1 reveals many common NLP applications where race has not been examined, such as machine translation, summarization, or question answering.[8] While some tasks seem inherently more relevant to social context than others (a claim we dispute in this work, particularly in §6.4), *research on race is compartmentalized to limited areas of NLP even in comparison with work on 'bias'*. For example, Blodgett et al. (2020) identify 20 papers that examine bias in co-reference resolution systems and 8 in machine translation, whereas we identify 0 papers in either that consider race. Instead, race is most often mentioned in NLP papers in the context of abusive language, and work on detecting or removing bias in NLP models has focused on word embeddings.

Overall, our survey identifies a need for the examination of race in a broader range of NLP tasks, the development of multi-dimensional data sets, and careful consideration of context and appropriateness of racial categories. In general, race is difficult to operationalize, but NLP researchers do not need to start from scratch, and can instead draw from relevant work in other fields.

6.4 NLP propagates marginalization of racialized people

While in §6.3 we primarily discuss race as a topic or a construct, in this section, we consider the role, or more pointedly, the absence, of traditionally underrepresented people in NLP research.

6.4.1 People create data

As discussed in §6.2.2, data and annotations are generated by people, and failure to consider who created data can lead to harms. In §6.2.2 we identify a need for diverse training data in order to ensure models work for a diverse set of people, and in §6.3 we describe a similar need for diversity in

[8] We identified only 8 relevant papers on Text Generation, which focus on other areas including chat bots, GPT-2/3, humor generation, and story generation.

data that is used to assess algorithmic fairness. However, gathering this type of data without consideration of the people who generated it can introduce privacy violations and risks of demographic profiling.

As an example, in 2019, partially in response to research showing that facial recognition algorithms perform worse on darker-skinned than lighter-skinned people (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019), researchers at IBM created the "Diversity in Faces" data set, which consists of 1 million photos sampled from the the publicly available YFCC-100M data set and annotated with "craniofacial distances, areas and ratios, facial symmetry and contrast, skin color, age and gender predictions" (Merler et al., 2019). While this data set aimed to improve the fairness of facial recognition technology, it included photos collected from a Flickr, a photo-sharing website whose users did not explicitly consent for this use of their photos. Some of these users filed a lawsuit against IBM, in part for "subjecting them to increased surveillance, stalking, identity theft, and other invasions of privacy and fraud."[9] NLP researchers could easily repeat this incident, for example, by using demographic profiling of social media users to create more diverse data sets. While obtaining diverse, representative, real-world data sets is important for building models, data must be collected with consideration for the people who generated it, such as obtaining informed consent, setting limits of uses, and preserving privacy, as well as recognizing that some communities may not want their data used for NLP at all (Paullada, 2020).

6.4.2 People build models

Research is additionally carried out by people who determine what projects to pursue and how to approach them. While statistics on ACL conferences and publications have focused on geographic representation rather than race, they do highlight under-representation. Out of $2,695$ author affiliations associated with papers in the ACL Anthology for 5 major conferences held in 2018, only 5 (0.2%) were from Africa, compared with $1,114$ from North America (41.3%).[10] Statistics published for 2017 conference attendees and ACL fellows similarly reveal a much higher percentage of people from "North, Central and South America" (55% attendees / 74% fellows) than from "Europe, Middle East and Africa" (19%/13%) or "Asia-Pacific" (23%/13%).[11] These broad regional categories likely mask further under-representation, e.g. percentage of attendees and fellows from Africa as compared to Europe. According to an NSF report that includes racial statistics rather than nationality, 14% of doctorate degrees in Computer Science awarded by U.S. institutions to U.S. citizens and permanent residents were awarded to Asian students, $< 4\%$ to Black or African American students, and 0% to American Indian or Alaska Native students (National Center for Science and Engineering Statistics, 2019).[12]

It is difficult to envision reducing or eliminating racial differences in NLP systems without changes in the researchers building these systems. One theory that exemplifies this challenge is *interest*

[9] https://www.classaction.org/news/class-action-accuses-ibm-of-flagrant-violations-of-illinois-biometric-privacy-law-to-develop-facial-recognition-tech#embedded-document
https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921 IBM has since removed the "Diversity in Faces" data set as well as their "Detect Faces" public API and stopped their use of and research on facial recognition. https://qz.com/1866848/why-ibm-abandoned-its-facial-recognition-program/

[10] http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/

[11] https://www.aclweb.org/portal/content/acl-diversity-statistics

[12] Results exclude respondents who did not report race or ethnicity or were Native Hawaiian or Other Pacific Islander.

convergence, which suggests that people in positions of power only take action against systematic problems like racism when it also advances their own interests (Bell Jr, 1980). Ogbonnaya-Ogburu et al. (2020) identify instances of interest convergence in the HCI community, primarily in diversity initiatives that benefit institutions' images rather than underrepresented people. In a research setting, interest convergence can encourage studies of incremental and surface-level biases while discouraging research that might be perceived as controversial and force fundamental changes in the field.

Demographic statistics are not sufficient for avoiding pitfalls like interest convergence, as they fail to capture the lived experiences of researchers. Ogbonnaya-Ogburu et al. (2020) provide several examples of challenges that non-white HCI researchers have faced, including the invisible labor of representing 'diversity', everyday microaggressions, and altering their research directions in accordance with their advisors' interests. Rankin and Thomas (2019) further discuss how research conducted by people of different races is perceived differently: "Black women in academia who conduct research about the intersections of race, gender, class, and so on are perceived as 'doing service,' whereas white colleagues who conduct the same research are perceived as doing cutting-edge research that demands attention and recognition." While we draw examples about race from HCI in the absence of published work on these topics in NLP, the lack of linguistic diversity in NLP research similarly demonstrates how representation does not necessarily imply inclusion. Although researchers from various parts of the world (Asia, in particular) do have some numerical representation among ACL authors, attendees, and fellows, NLP research overwhelmingly favors a small set of languages, with a heavy skew towards European languages (Joshi et al., 2020) and 'standard' language varieties (Kumar et al., 2021).

6.4.3 People use models

Finally, NLP research produces technology that is used by people, and even work without direct applications is typically intended for incorporation into application-based systems. With the recognition that technology ultimately affects people, researchers on ethics in NLP have increasingly called for considerations of whom technology might harm and suggested that there are some NLP technologies that should not be built at all. In the context of perpetuating racism, examples include criticism of tools for predicting demographic information (Tatman, 2020) and automatic prison term prediction (Leins et al., 2020), motivated by the history of using technology to police racial minorities and related criticism in other fields (Browne, 2015; Buolamwini and Gebru, 2018; McIlwain, 2019). In cases where potential harms are less direct, they are often unaddressed entirely. For example, while low-resource NLP is a large area of research, a paper on machine translation of white American and European languages is unlikely to discuss how continual model improvements in these settings increase technological inequality. Little work on low-resource NLP has focused on the realities of structural racism or differences in lived experience and how they might affect the way technology should be designed.

Detection of abusive language offers an informative case study on the danger of failing to consider people affected by technology. Work on abusive language often aims to detect racism for content moderation (Waseem and Hovy, 2016). However, more recent work has show that existing hate speech classifiers are likely to falsely label text containing identity terms like 'black' or text containing linguistic markers of AAE as toxic (Dixon et al., 2018; Sap et al., 2019; Davidson et al., 2019; Xia et al., 2020). Deploying these models could censor the posts of the very people they purport to help.

In other areas of statistics and machine learning, focus on *participatory design* has sought to amplify the voices of people affected by technology and its development. An ICML 2020 workshop titled "Participatory Approaches to Machine Learning" highlights a number of papers in this area (Kulynych et al., 2020; Brown et al., 2019). A few related examples exist in NLP, e.g. Gupta et al. (2020) gather data for an interactive dialogue agent intended to provide more accessible information about heart failure to Hispanic/Latinx and African American patients. The authors engage with healthcare providers and doctors, though they leave focal groups with patients for future work. While NLP researchers may not be best situated to examine how people interact with deployed technology, they could instead draw motivation from fields that have stronger histories of participatory design, such as HCI. However, we did not identify citing participatory design studies conducted by others as common practice in the work we surveyed. As in the case of researcher demographics, participatory design is not an end-all solution. Sloane et al. (2020) provide a discussion of how participatory design can collapse to 'participation-washing' and how such work must be context-specific, long-term, and genuine.

6.5 Conclusions and Future Work

We conclude by synthesizing some of the observations made in the preceding sections into more actionable items. First, NLP research needs to explicitly incorporate race. We quote Benjamin (2019): *"[technical systems and social codes] operate within powerful systems of meaning that render some things visible, others invisible, and create a vast array of distortions and dangers."*

In the context of NLP research, this philosophy implies that all technology we build works in service of some ideas or relations, either by upholding them or dismantling them. Any research that is not actively combating prevalent social systems like racism risks perpetuating or exacerbating them. Our work identifies several ways in which NLP research upholds racism:

- Systems contain representational harms and performance gaps throughout NLP pipelines
- Research on race is restricted to a narrow subset of tasks and definitions of race, which can mask harms and falsely reify race as 'natural'
- Traditionally underrepresented people are excluded from the research process, both as consumers and producers of technology

None of these challenges can be addressed without direct engagement with marginalized populations. NLP researchers can draw on precedents for this type of engagement from other fields, such as participatory design and value sensitive design models (Friedman et al., 2013). Additionally, numerous organizations already exist that serve as starting points for partnerships, such as Black in AI, Masakhane, Data for Black Lives, and the Algorithmic Justice League.

Nevertheless, race and language are complicated, and while readers may look for clearer recommendations, no one data set, model, or set of guidelines can 'solve' racism in NLP. For instance, while we draw from linguistics, Hudley et al. (2020) in turn call on linguists to draw models of racial justice from anthropology, sociology, and psychology. Relatedly, there are numerous racialized effects that NLP research can have that we do not address in this work; for example, Bender et al. (2021) and Strubell et al. (2019) discuss the environmental costs of training large language models, and how global warming disproportionately affects marginalized communities.

Finally, while this chapter focuses on race, which we note has received substantially less attention than gender, many of the observations in this work hold for social characteristics that have received even less attention in NLP research, such as socioeconomic class, disability, or sexual orientation (Mendelsohn et al., 2020; Hutchinson et al., 2020). Our work serves to highlight some of the risks and harms that can result from NLP technology by focusing on a single social dimension, and we suggest that readers consider our work as one starting point for bringing inclusion and justice into NLP.

Chapter 7

Ethical Implications

The rapid improvement and expansion of language technologies has led to their increasing integration into systems that can have substantial impacts on people's lives. The potential positive and negative impacts are particularly pronounced in language technologies developed to addressed social issues that process data relevant to current events. This section provides discussion of the ethical implications considered throughout this work.

Dual Use Much of the methodology and frameworks developed in this work have the potential to be misused, resulting in dual-use problems (Hovy and Spruit, 2016). Methodology to uncover stereotypes could be used to reinforce them (§2); characterizations of information manipulation campaigns can provide strategies to spreaders of disinformation (§3); identification of online harassment can be used to aid in promoting and imitating it (§4). We persist in this work despite the possibility of misuse under the belief that the phenomena studied in this work are already in existence, and studying and publicizing them is essential for mitigating them. For example, while exposing propaganda strategies could be construed as writing a guide for generating propaganda, publishing known strategies and increasing public knowledge of manipulation campaigns is likely to mitigate their influence on public opinion. Nevertheless, it is essential to continuously review and critique the risks and benefits of future research in these areas.

Data and Privacy This work involves the use of a wide variety of data, including news articles, social media posts, Wikipedia articles, and child welfare case notes. The child welfare case notes (§5.2) contain highly sensitive information about people who did not explicitly consent to this research. All research was conducted only after a full IRB review, and all data was stored on a secure disk-encrypted server with restricted access. All researchers additionally aimed to avoid reading notes as much as possible and in order to preserve privacy of the people described in them. Throughout §5.2 we do not include any identifying information from the data, and any references to specific content are only shown in aggregated or are paraphrased at an extremely high level.

Additionally, while all of the social media data considered in this work (§4, §3.3, §5.1) was publicly available at the time of data collection, posts were made by users who did not explicitly consent to this analysis. In general, we do not identify any individual social media users, nor make any attempt to predict characteristics about private citizens, and when referencing examples from sensitive posts, we paraphrase content. Additionally, in accordance with platform terms of service, we do not publicly release any raw text data, preserving users' ability to delete content. Vitak et al. (2016) and Williams et al. (2017) provide a more in-depth discussion of ethical considerations of research using social media data.

In some cases, we do provide discussion and examples of individuals, primarily examining how they are described in news articles or on Wikipedia (§2). This work is exclusively focused on how individuals are described and not necessarily reflective of their true actions or characteristics. In these cases, we focus on analysis of public figures, including politicians, athletes, and celebrities, where there is a lower expectation of privacy. Similarly, in §4.2 we do train a model to predict gender, but this model relies on self-identified gender of public figures and is designed to identify comments addressed towards people of different genders. It is not usable for attempts to infer gender automatically. Nevertheless, despite our focus on how people are described or addressed in text, this work does involve depictions of real people, which could have unforeseen consequences on them and their public images.

Simplification and Categorization The problems addressed in this work are complex and multifaceted and approaching them computationally requires assumptions and simplifications. One concept exemplifying this type of simplification is the use of categorizations schema (e.g. gender and race, §2, §4.2, §5.2) throughout this work (Hanna et al., 2020). As much as possible, we strive to be intentional and explicit about these schema and the contexts they are derived from, but we clarify here that they are simplifications of complex social characteristics.

Similarly, we avoid and discourage any framing of this work that consider the methods developed here or in related research as "solutions". Concepts like propaganda, online toxicity, and stereotyping are complex social issues. Similarly, the child welfare system is deeply complex. For example, even perfect technology for aiding caseworkers in providing services is insufficient for addressing the underlying issues like poverty and addiction that prompt involvement of government agencies in child welfare and for addressing ways that the current system can colossally fail families (Eubanks, 2018; Abebe et al., 2020). We persist in this work on the belief that technology can be useful in these settings, even though it is insufficient on its own and in some cases serves as incremental temporary improvement (Green, 2019). We further aim to maximize the potential utility of this work through regular consultation and collaboration with domain experts, including researchers at the Wikimedia foundation (§2.3), the Allegheny Department of Human Services (§5.2), and non-profit organizations (Data for Black Lives, §5.1). Throughout this work (e.g., §2.3, §5.2) we generally suggest that this research aims to augment and support human analysis, intervention, and decision-making and cannot replace them.

Power Imbalances and Stakeholder Participation We further acknowledge that this work was conducted primarily at an academic institution and reflects power imbalances common in NLP research, in that technologists have the power to decide which projects to pursue, even though much broader communities may be affected by those decisions (Blodgett et al., 2020). One of the motivating beliefs behind this work is that developing NLP technology to reducing stereotyping, manipulation, and disparate performance can aid in empowering underrepresented people. Nevertheless, while we do engage with domain experts, in most of this work we do not directly engage with all relevant stakeholders (e.g., we engage with the Department of Human Services, but not families receiving services). Direct engagement does impose work on participants (Sloane et al., 2020), and we take the view that drawing from prior studies (Brown et al., 2019) and engaging with domain experts provides some guidance to academic research without imposing additional burdens. Nevertheless, we caution that none of the technology developed in this work is intended to be off-the-shelf deployable, and we do not condone deployment without further investigation of potential impacts and engagement with community stakeholders likely to be most affected.

Potential Impact of Deployment From the research presented in this work, there are observable impacts that could result from the deployment of this technology. None of the methodology developed or explored has perfect accuracy. In §2, where the proposed methodology is most useful for aiding in writing and editing content or balancing training data, false negatives are the main concern: errors in measuring portrayals of people could miss biased content and false certify content as "unbiased". Thus, we note that these methods cannot be used to certify data quality on their own. These methods could also unintentionally influence journalists and editors to write biased content, for example, encouraging them to portray women with higher power, even if that portrayal is not reflective of reality.

False negatives and false positives are both problematic in identifying propaganda strategies (§3) and toxicity (§4), as false positives can result in unwarranted censorship and false negatives can miss harmful content. Even if models were perfect, this work could result in unintended censorship, especially given the subject nature of the content studied, e.g. polarizing language can be considered a legitimate form of political speech, and discouraging or demoting it could result in marginalization. Additionally, this work carries the risk of promoting toxic or manipulative content that may otherwise have gone unnoticed.

Models for emotion detection risk misconstruing content and discussing characteristics of protest movements can result in highlighting ways to derail them or increasing violence (§5.1). Ekman's work in particular has been heavily criticized (Crawford et al., 2021), though our use case focuses on measuring emotions intentionally expressed in public, and not ones that people may not wish to share. §5.2 and §6 highlight additional specific risks of deploying NLP models, including the potential to reinforce preconceived or historical biases, and Chouldechova et al. (2018) and Brown et al. (2019) provide additional discussion on the risks of deploying AI in general in child welfare settings.

Self-disclosure As a PhD student at a U.S. institution, I am situated within the traditionally exclusionary practices of academic research. This perspective has impacted my work, and there are viewpoints outside of my experience that this work may not fully represent.

Chapter 8

Conclusions

8.1 Summary of Contributions

- I introduce *contextual affective analysis*, a framework grounded in social psychology for analyzing how people are portrayed in narrative text, and I present *verb-centric* and *entity-centric* methodology for implementing it.

- I release data sets to facilitate additional research on portrayals of people: online media articles about the #MeToo movement and Wikipedia biography pages with inferred attributes.

- I develop a matching algorithm to facilitate controlled analyses of articles with sparse attributes. I demonstrate its usefulness in analyzing Wikipedia biography pages, revealing how prior work has been unable to control for confounding variables.

- I introduce identifying framing and agenda setting as tasks for NLP detection of opinion manipulation strategies. I present case studies examining these concepts in Russian news as well as polarization in Indian and Pakistani social media.

- I propose approaching toxicity detection as an unsupervised task. In developing an unsupervised approach, I demonstrate how integrating causal frameworks into NLP can encourage models to learn deeper pragmatic features.

- I conduct an examination of emotions expressed in tweets about #BlackLivesMatter protests, which shows how NLP can provide information to community organizers and debunk stereotypes when adapting to specific domains.

- I empirically analyze the risks and benefits of integrating text features in a risk assessment tool used in processing child welfare referrals. This work highlights ethical concerns in NLP research and has the potential to influence policy on child welfare cases.

- I conduct a survey of NLP literature on race, highlighting limitations in current work and how research from related fields can improve NLP approaches.

8.2 Discussion and Future Work

This thesis presents methodology and frameworks for developing NLP technology to address complex social issues. While this thesis is organized by application domain, several common themes arise across chapters, which I highlight in this section and discuss as avenues for future work.

Domain-specific NLP This thesis emphasizes that social issues and associated text are extremely context-dependent and vary greatly in different domains. Child welfare case notes use different vocabulary and acronyms than domains like news (§5.2.2), protest movements subvert traditional word connotations (§5.1), and concepts like "gender bias" manifest different in different contexts (§4.2). Not only is data domain-specific, in practical settings, it is often much messier than data sets curated for training and evaluating models, containing acronyms, typos, and inconsistent formatting (§5.2.2).

While domain-adaption has long been an area of NLP research, recent years in particular have seen much focus on general-purpose NLP models that are pre-trained on large amounts of text data and usable in a broad range of NLP tasks (e.g. (Devlin et al., 2019; Liu et al., 2019)). These types of models have been shown to adapt well to new tasks through few-shot learning (Brown et al., 2020), but this thesis suggests that they can be unusable in social-oriented applications due to the need for domain-specificity. In §5.2.2, we do not use a pre-trained language model due to domain-specific vocabulary and the need for large amounts of annotated data to sufficiently update the model. In §5.1, we do use a pre-trained language, but only with additional pre-training and fine-tuning on in-domain data. Both sections suggest that the most domain-specific language is also often then most informative.

For NLP models to be usable in practice and particularly in emerging scenarios where connotations and language can change rapidly, there is need to develop better methods for adapting models trained on out-of-domain data without requiring many in-domain annotations. Even small amounts of in-domain annotations can time-consuming to collect and require multiple rounds of annotation and scheme revision, as exemplified in §A.1.2.

Reducing Supervision Relatedly, not only is collecting annotated data time-consuming and expensive, it is often infeasible for concepts that are subtle and hard-to-define. For example, Sap et al. (2019) show how crowd-sourced annotations of toxicity can exhibit racial bias, and Breitfeller et al. (2019) use the assumption that there are discrepancies in perceived offensiveness of microaggressions between dominant groups and marginalized groups in order to build a corpus of annotated microaggressions by focusing on instances where annotators disagreed. This work motivates our development of an unsupervised approach to identifying toxicity in §4.2, and we similarly find that gender bias differs in different domains, making it difficult to annotate in isolated instances.

In §4.2, while our method is unsupervised in that it does not rely on "bias" annotations, we do train a supervised model, where the prediction task is the gender of the addressee. This approach has some similarity to our methodology for identifying agenda-setting in §3.1, which compares text content with Russian economic indicators, and also with the AFST risk assessment tool that we examine in §5.2. These approaches rely on using existing external variables to train NLP models and classify or analyze text. In §4.2 and §3.1, these variables do reflect exactly the content we are aiming to learning: text predictive of gender and correlations with economic indicators. However, in many circumstances, tasks rely on proxy variables, as is the case in §5.2, where the AFST uses future placement as a proxy for risk (Coston et al., 2020). Thus, external variables are often not a reliable source of supervision.

Furthermore, much of the NLP research that has proven useful in other settings, particularly in social science research, has been unsupervised methods for text analysis. Topic modeling (Blei et al., 2003; Roberts et al., 2013) remains a popular tool for examining new data sets, and more recent work has sought to use word embeddings for similar analyses (Garg et al., 2020; Joseph and Morgan, 2020). Continued research is needed to develop unsupervised methods for text processing, including understanding how advances in self-supervised training objectives can be incorporated in analyses of specific corpora or short text samples.

Causality and Confounding Variables The use of external indicators in §3.1, §4.2, and §5.2 more broadly relates to the need for incorporation of social context in NLP models. Language use and desired behavior of NLP models can vary greatly even within the same domain, depending on characteristics of people writing text or using models. In some cases, contextualizing information is important for dictating expected model behavior, e.g., a dialog agent in a workplace setting is expected to use more formal language than on social media. In much of this thesis, context including dialect (§4.1), characteristics of people (§2.3, §4.2), and events (§5.2) can become confounds whose influence needs to be reduced in NLP models and text analysis.

NLP models are commonly developed over benchmark tasks and data sets (e.g. (Wang et al., 2019)). Focusing on performance metrics is useful for model development; however, models can often obtain high performance by learning factors correlated with the target task, rather than actually learning the target task (Kaushik and Lipton, 2018). In some settings this type of performance is not problematic: models may fail to generalize to other types of data, but may perform sufficiently well to be deployed in specific settings. In most social-oriented tasks, and particularly in social science analyses of concepts like stereotypes and bias in text, measuring the target values of interest is critical.

In general there is rich literature in causal inference that can provide guidance. §2.3 draws from this literature in introducing a matching-based approach to controlling for confounding variables, while §4.1 uses adversarial training and §4.2 incorporates both. These methods can serve as starting points, but they do have limitations and there is need for more work in this area, including developing both methodology and evaluation criteria.

Accountability, Transparency, and Ethics Finally, greater use of NLP in addressing social issues raises challenges around who is accountable for system performance and failures, how to ensure use of algorithms is transparent to people affected by them, and ethical issues around data collection, model development, and technology deployment. §7 discusses the concerns specific to this thesis in more depth and §6 discusses some of the broader implications of NLP research through one particular lens.

The increasing deployment of NLP and its usability in social-oriented tasks is still a very new area of research, and much work understanding its practical effects on society remains to be done. Many of the tasks themselves discussed in this work are new, as toxicity on social media, widespread opinion manipulation, and online stereotyping are 21st century issues. Further understanding of risks and benefits of NLP, including continued critique of what NLP research should and should not be pursued will be essential to minimizing the potential harms of this work.

Appendix A

Appendix

A.1 Data and Methodology Details for #BlackLivesMatter Analysis

A.1.1 List of Search Terms for Data collection and splits

Below is the list of words and hashtags used to retrieve the data from Twitter's Standard search API, which at the time of data collection allowed for retrieving tweets by hashtags and keywords for up to 7 days prior. We group them as follows when dividing tweets by keywords:
Pro-BLM Hashtags #BlackLivesMatter, #GeorgeFloyd, #ICantBreathe, #BLM, #JusticeForFloyd, #JusticeForGeorgeFloyd, #GeorgeFloydProtests, #WorldAgainstRacism, #WalkWithUs, #KneelWithUs, #BlackoutTuesday, #VoteOutHate, #NoJusticeNoPeace, #BlackWomenMatter, #BlackGirlsMatter, #TheShowMustBePaused
Anti-BLM Hashtags #BlueLivesMatter, #AllLivesMatter, #AllLivesMatters, #AllLivesMattter, #WhiteLivesMatter, #WhiteLivesMatters, #WhiteLivesMatterMost, #WhiteLivesMatterMore, #WhiteLifeMatters, #WhiteLifeMatter
Police cops, police,
Protests protests, protesters, protestors,
Other george floyd, derek chauvin, protest, riot, riots, rioters, looting, looters,

A.1.2 Materials and methods

Model setup Our primary classifier for identifying emotions uses pre-trained BERT as the base network. In emotion classification, BERT has consistently outperformed other models such as convolutional neural networks and RoBERTa (Desai et al., 2020; Devlin et al., 2019; Liu et al., 2019). We append a 2-layer feedforward neural network on top of BERT, which takes the mean-pooled representation of all input tokens. We train one classifier per emotion which makes each task a binary classification task. We also experimented with multi-class classification, but we found little difference in performance and ultimately use single-class models to ensure that any identified correlations between emotions are not model artifacts. We used the `bert-base` model from the transformers library (Wolf et al., 2020) with the same hyperparameters as the ones initially implemented for BERT. The mean-pooled embedding from BERT with a dimension of 768 goes through a dropout

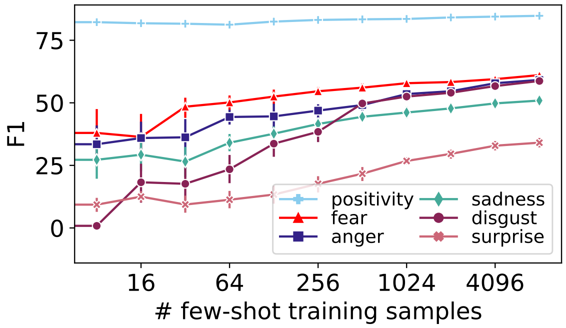


Figure A.1: F1 scores of emotion classifiers on HurricaneEmo test data, using GoEmotions and varying number of few-shot training samples from HurricaneEmo as training data. Results are averaged across 10 random seeds, and the error bar indicates the 95% confidence interval (CI).

layer with a dropout probability of 0.1 and then goes into the final classifier layer that consists of a 2-layer feedforward neural network. The hidden size of the classifier is set to 256, and the dropout probability is set to 0.5. The total number of trainable parameter is 110M. We did not perform any hyperparameter tuning and used the same settings for all models throughout the experiments.

We used batch size of 32, and used AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-5 to train the models. The weight decay coefficient was set to 0.01, and gradients were clipped to ($[-5, 5]$). Keung et al. (2020) report that source performance on validation set is often uncorrelated with target validation performance, and suggest using the target validation set for model selection even in the zero-shot setting. Following this suggestion, we used the validation split of HurricaneEmo to choose the final model in both zero-shot and few-shot learning settings. For the zero-shot models, we trained for minimum 2,500 steps and trained until the early stopping criteria was met (patience of 3). When there was additional few-shot training, we trained for another minimum 500 steps and trained until the same stopping criteria was reached. Training was done on a single-GPU (NVIDIA RTX 2080 Ti) machine, and it took about 20 minutes to run one zero-shot experiment, and additional 10 minutes to run few-shot learning after it. We ran 10 experiments per model with different random seeds (15, 17, 19, 23, 29, 31, 33, 37, 41, 43) and reported averaged scores across seeds.

Few-shot Training Data Size In order to finalize the +FSL model, we use GoEmotions with small subsets of HurricaneEmo as training data and HurricaneEmo as test data to experiment with different in-domain data set sizes. Figure A.1 reports results. Unsurprisingly, performance improves as the size of the in-domain training data increases. However, the rate of improvement is not standard for all emotions. Prediction of POSITIVITY changes little with increasing data set sizes, while prediction of DISGUST shows the greatest changes. The steepest rate of improvement occurs between 0 and 256 data points, after which we see diminishing returns for most emotions. Based on these results, we fix the in-domain data size for the +FSL models to 300 and annotated 700 instances from #BLM2020 to facilitate few-shot learning and evaluation.

Plutchik-Ekman mapping The 8 Plutchik emotions map directly to the 6 Ekman emotions except for 2 additional categories. The first category, ACCEPTANCE, includes sub-emotions of "acceptance", "trust", and "admiration". Given Ekman's definition of JOY as all positive emotions and the suggested mapping of "adminration" to JOY for the GoEmotions data (Demszky et al., 2020), we map these 3 emotions to POSITIVITY. The second category, EXPECTANCY, includes "vigilance", "interest", and "anticipation". Based on Ekman's definitions and the inclusion of "nervousness" in "fear" under the GoEmotions mapping, we map "interest" and "anticipation" to POSITIVITY, and we match "vigilance" to FEAR. Thus, we impose the same Ekman-style mapping on the Hurricane data set as on the GoEmotions data set to evaluate the model across data sets.

Data Pre-processing We take the following pre-processing steps over all data: (1) replacing URLs, user references, and numbers with special tokens; (2) replacing emojis with their text descriptions, e.g., ☺ → "<e> smiley face </e>"[1]; (3) truncating repetitions of punctuation and user reference to at most three occurrences; (4) segmenting the words in hashtags, e.g. *#BlackLivesMatter*→ <*h*> *black lives matter* </*h*>[2]; (5) lowercasing, but keeping upper case information by prepending a special token, e.g., *SHOUT*→ <all_caps> *shout*.

#BLM2020 Annotations We collected an initial set of annotations over 400 tweets from #BLM2020, which were conducted by 5 volunteers who were living in the United States throughout the time period in our data set. In the annotation instructions, annotators were provided with all sub-emotions used in GoEmotions for each high-level emotion (listed in §5.1.2) and asked to select all the emotions that occurred in the tweet, either expressed by the author or solicited in the reader. Annotations were conducted by 5 volunteers who were living in the United States throughout the time period in our data set. For each tweet, we collected two independent judgments. If the two annotators disagreed on any label, a third independent annotation was collected. Final labels for each tweet were obtained using majority-voting. We ultimately use these initial annotations as training (~ 300) and development sets (~ 100).

In order to ensure annotation quality in our test set, we revised the annotation scheme based on feedback from the initial annotations and collected annotations over an additional 300 tweets. Notably, the revised schema allowed annotators to distinguish between emotions that were *clearly* expressed in the tweet and *maybe* expressed in the tweet. In addition to the written instructions, we had a 45min training call with all annotators. During the call, we asked annotators to independently annotate 10 tweets to ensure that they understood the instructions and were able to use the annotation interface. The second round of annotations was conducted by 6 total annotators, where all annotators annotated every tweet. Annotators were offered compensation of $150, though some annotators declined. Final labels were obtained by weighting *clearly expressed* annotations as 1 and *maybe expressed* annotations as 0.5, and a tweet was considered to contain an emotion if the total sum of weighted annotations was > 3. We note that only one annotator annotated tweets from both the training/development and test set. In keeping annotator judgements primarily separate, we aim to avoid modeling judgements of specific annotators (Geva et al., 2019). In both rounds of annotations, we additionally asked annotators to label tweets as supportive of the Black Lives Matter movement, opposed to the movement, related to the movement but unable to tell stance, or unrelated. In aggregating these annotations, if $> \frac{1}{2}$ labeled the tweet with the same stance we assign the tweet that label, otherwise we assign it as unable to tell.

[1] https://pypi.org/project/emoji/

[2] https://pypi.org/project/wordsegment/

Label	α_{train}	α_{test}	$Corr_{train}$	$Corr_{test}$
ANGER	0.38	0.49	0.74	0.72
DISGUST	0.43	0.55	0.72	0.76
POSITIVITY	0.66	0.71	0.83	0.74
SURPRISE	0.25	0.26	0.65	0.54
SADNESS	0.43	0.35	0.55	0.70
FEAR	0.36	0.50	0.64	0.51
BLM-stance	0.61	0.55		

Table A.1: Inter-annotator agreement (Krippendorff's α) and interrater correlation (Delgado and Tibau, 2019) over 400 train/dev and 300 test tweets for emotions and stance in our data set of tweets about #BlackLivesMatter.

Table A.1 reports inter-annotator agreement specifically Krippendorff's α, which is a commonly used metric, and interrater correlation (Delgado and Tibau, 2019), which is a metric used in prior work (Demszky et al., 2020). In computing Krippendorff's α over the test set, we only consider annotators to disagree if one selected *clearly expressed* and another did not identify an emotion to be expresssed at all (e.g. *clearly expressed* and *maybe expressed* annotations are considered agreement). In computing interrater correlation, which requires quantification of annotations, we use the same 0.5/1 weighting as we use when aggregating annotations. Given the subjective nature of emotions and that we avoid directing annotators on how to annotate particular types of tweets to avoid unduly influencing results, some disagreement is expected over this data. Agreement over our data set is generally higher than the agreement reported for GoEmotions (Demszky et al., 2020). Test agreement is additionally high for the emotions we focus on in our analysis POSITIVITY, ANGER, and DISGUST. While agreement over ANGER and DISGUST is lower than agreement over POSITIVITY, most disagreements occurred in annotators struggling to distinguish ANGER and DISGUST from each other (Salerno and Peter-Hagene, 2013). When we collapse ANGER and DISGUST, agreement rises to (0.57/0.79) over the training set and (0.68/0.80) over the test set.

Krippendorff's α over the test set is lower over SURPRISE and SADNESS than other emotions. We suspect lower agreement over SADNESS relates to the sparsity of this emotion in our test set and note that interrater correlation is comparable to other emotions. Discussions with annotators suggested that SURPRISE agreement is low due to the rareness of genuine expressions of surprise or confusion and differing interpretations of devices like rhetorical questions and sarcasm, which some annotators marked as SURPRISE and some did not. Given the lower agreement, while we include measurements of SURPRISE for completeness, we avoid drawing conclusions about this emotion category.

A.1.3 Detailed Model Performance

Table A.2 shows the precision, recall, and F1 scores for all models when training on GoEmotions+HurricaneEmo and evaluating on #BLM2020 (the same setup as Figure 5.2).

As the annotated #BLM2020 consists of only a few hundred data points, which is too small to provide a comprehensive evaluation, we additionally validate our model on the test partition of HurricaneEmo. More specifically, we use the train partition of GoEmotions as training data, small subsets of the HurricaneEmo train partition for few-shot learning, and the HurricaneEmo test partition for evaluation. Tweets about hurricanes involve strong emotions in a domain-specific setting, where the central event is inherently negative. Thus, they have similarities to our eventual

	anger	**disgust**	**fear**	**positivity**	**sadness**	**surprise**
LIWC	**68.6**/54.7/60.9	-	-	30.0/37.5/33.3	25.0/40.0/30.8	-
BASE	64.7/68.0/66.2	83.9/17.0/27.9	25.4/23.1/24.0	32.8/85.9/47.3	35.6/38.1/36.3	20.1/44.7/27.8
+TGT	69.3/64.2/66.6	**84.9**/25.4/38.5	26.7/22.0/23.9	32.0/**89.2**/47.3	**38.1**/**45.2**/**41.2**	17.9/**46.1**/25.5
+FSL	60.2/84.7/70.4	78.2/57.6/66.2	32.3/28.6/28.8	**45.9**/72.3/55.8	36.9/23.5/28.3	**42.0**/26.8/**32.3**
+TGT+FSL	64.3/**85.0**/**73.1**	81.8/**64.8**/**72.2**	**41.3**/**36.6**/**37.5**	45.3/81.8/**58.2**	35.9/22.0/26.5	36.8/27.4/31.2

Table A.2: **HurricaneEmo+GoEmotions→#BLM2020.** Precision/Recall/F1 scores of models trained with GoEmotions and HurricaneEmo and evaluated on the test set of the annotated samples from #BLM2020. The best score in each emotion and in each metric is boldfaced.

	anger	**disgust**	**fear**	**positivity**	**sadness**	**surprise**
LIWC	28.1/26.0/27.0	-	-	76.3/37.7/50.5	35.3/20.6/26.0	-
BASE	40.8/37.2/38.5	17.0/ 2.0/ 3.3	**68.4**/11.4/19.4	78.4/80.8/79.6	37.6/34.4/35.8	13.0/21.6/16.1
+TGT	49.0/39.0/43.5	27.2/ 6.8/10.2	62.2/35.6/44.7	78.2/83.8/81.0	46.0/31.4/37.2	9.2/**48.4**/15.6
+FSL	45.4/35.0/39.1	47.8/34.0/39.3	51.2/**61.4**/55.7	78.2/**87.4**/82.4	44.6/39.8/42.0	28.4/17.0/**20.9**
+TGT+FSL	**56.0**/**39.2**/**45.9**	**49.6**/**40.6**/**44.6**	56.4/55.0/**55.8**	**79.4**/86.8/**83.0**	**51.2**/**42.4**/**46.6**	**28.6**/15.0/19.4

Table A.3: **GoEmotions→HurricaneEmo.** Precision/Recall/F1 scores of models trained with GoEmotions and evaluated on HurricaneEmo test set. The best score in each emotion and in each metric is boldfaced.

analysis corpora and serve as a meaningful evaluation set for model performance. Table A.3 presents the precision, recall, and F1 scores for all models in this setting.

In both Table A.2 and Table A.3, we see that +TGT+FSL performs generally well in all three metrics, and the performance gain of +TGT+FSL compared to the baseline (BASE) often comes from the improved precision.

A.1.4 User location inference

Identifying user locations We identify locations for users in our data set based on the user-populated location string in their profile, which is typically populated more often than geolocation (Hecht et al., 2011; Alex et al., 2016) and was non-empty for 62.36% of users in our data set. We assigned state-level locations based on explicit mentions of U.S. states. We additionally identified mentions of cities listed as having protests in the ACLED data. For every user string that occurred at least 100 times and mentioned a city listed in the ACLED data, we manually resolved if the user string referred to a U.S. city and state (e.g. identifying "Los Angeles" as indicating California). We additionally examined the 100 most common user strings that we had not already affiliated with states in order to identify common acronyms, like "NYC" and "Philly". We ultimately identified 20.66% of users in our data set as affiliated with a U.S. state. We additionally mapped 12.3% of users to U.S. cities listed in the ACELD data. Some of the most common user strings that we did not map to U.S. states or cities include ones that referred to the U.S. broadly ("United States" listed by 0.94% of users; "USA" listed by 0.2% of users) or ones that did not refer to the United States ("London, England", specified by 0.55% of users; "Lagos, Nigeria" listed by 0.22% of users). We also do not include 0.04% (7,471) of users who specified multiple states.

Normalization of state-level protest data by county When we normalize by county count, in the ACLED data the states with the highest volumes of protests are: Washington D.C., Connecticut, Massachusetts, California, Delaware, and the states with the lowest volume of protests are: North Dakota, South Dakota, Mississippi, Nebraska, Arkansas. In the CCC data, the states with the highest volumes of protests are: Connecticut, Massachusetts, California, New York, New Jersey,

	Average	Min.	Max.
Number of tokens per note	156.59	0	2,915
Notes per case	128.36	1	4,831
Notes per referral	8.55	1	672

Table A.4: Overview statistics of 3.1M contact notes associated with cases and referrals.

and the states with the lowest volume of protests are: North Dakota, South Dakota, Mississippi, Nebraska, Kansas. Thus, this normalization shows more protests in traditionally liberal states and fewer protests in traditionally conservative states, and does not separate states purely by population or geographic size.

Normalization of city-level protest data by population In computing correlations against protest data at a city level, we weight protests by size estimates and normalize by city population counts from the 2020 U.S. census.[3] When the size of the protest is given as a range, we use the higher value, and when no protest size estimate is provided, we use the mean of size estimates for other protests in the same city. We found no differences in results when using the lower value of size ranges instead of the higer value. In the ACLED data, 38.2% $(2,579/6,741)$ of protest events are missing size estimates or had formats of size estimates we were unable to parse. In the CCC data, 38.6% $(3,256/8,420)$ protest events are missing size estimates. To reduce noise, we restrict analysis to cities with a population of $> 50,000$ and to cities were we associated > 500 users in our data set. These restrictions resulted in 337 cities from the ACLED data and 335 cities from the CCC data. Using this normalization, in the ACLED data the cities with the highest volumes of protests are: Camden (New Jersey), Santa Cruz (California), Federick (Maryland), Seattle (Washington), Salt Lake City (Utah) and the cities with the lowest volume of protests are: Apple Valley (Minnesota), Elyria (Ohio), Orem (Utah), Lawrence (Kansas), and Indio (California). In the CCC data, the cities with the highest volumes of protests are: District of Columbia, Oakland (California), Seattle (Washington), Providence (Rhode Island), Berkeley (California). The cities with the lowest volumes of protests are McAllen (Texas), Apple Valley (Minnesota), Suffolk (Virginia), Miami Beach (Florida), Temple (Texas).

A.2 Data and Methodology Details for Child Welfare Casenotes Analyses

A.2.1 Overview Statistics of contact notes

The full contact notes data set consists of $3,105,071$ notes. Table A.4 provides some overview statistics of this data, and Figure A.2 provides a histogram of the types of contacts these notes describe.

[3] https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-cities-and-towns-total.html

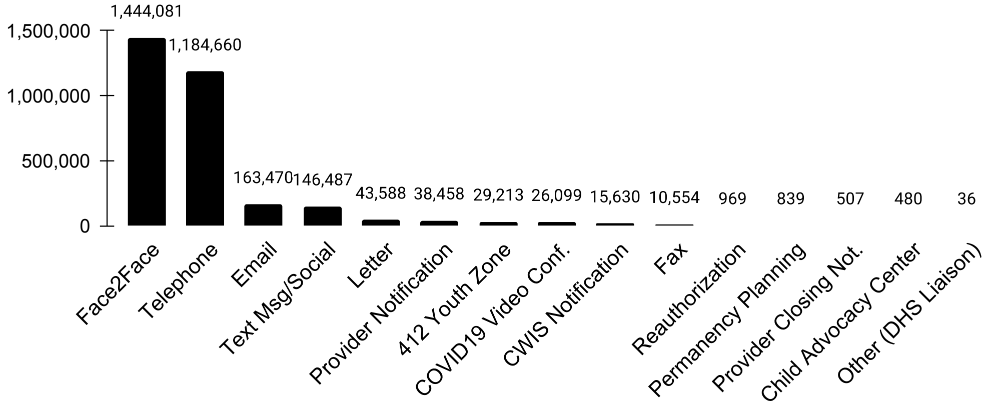


Figure A.2: Histogram over the type of contact recorded in each of 3.1M contact notes

Text Data Source	Total Train Size	Train size with text	Total Test Size	Test size with text
Previous Notes	28,769	7,893	14,417	4,133
Current Notes	28,769	28,340	14,417	14,417

Table A.5: Data set sizes for different prediction tasks. We follow prior work in only including referrals that were screened in (Chouldechova et al., 2018). *Previous Notes* refers to contact notes preceding the current referral (our primary data set). *Current Notes* refers to contact notes associated with the current referral.

A.2.2 Hyper-parameter Settings

In the RoBERTa model, we concatenate all associated notes and truncate them to the last (most recent) 512 tokens.[4]. We fine-tuned the model for classification using a learning rate of 1e-05 and weight decay of 0.01 for up to 30 epochs, where training was stopped early if development set performance did not increase for 3 epochs.

For the GatedCNN model, associated notes were truncated to the last (most recent) 3,000 tokens. The model embeddings were initialized with 100-dimensional word embeddings trained over the full data set of 3.1M contact notes using skip-gram Word2Vec with a context window of 5. We use the same kernel, filter sizes, and hidden layer sizes as the original model (Ji et al., 2021). The model was trained with a learning rate of 1e-03 for up to 20 epochs, with early stopping if development set performance did not increase for 3 epochs.

Hyper-parameters were selected based on development set performance after 5 epochs of training.

A.2.3 Additional data statistics and metrics

[4] In early experiments, we found that truncated outperformed alternative approaches to handling long inputs to a transformer, such as hierarchical models or selecting inputs using scoring functions

Description
Juveniles who are active with the Juvenile Probation Office (JPO) through supervision, placement or other services.
Clients receiving behavioral health inpatient help.
Defendants in cases processed by Allegheny County Courts.
Clients receiving behavioral health outpatient help.
Individuals and families receiving prevention services, support services and/or housing who are homeless or at risk of becoming homeless. Services are provided by DHS and DHS-contracted providers and include housing assistance, case management, prevention and outreach.
Individuals receiving a publicly-funded (Allegheny County or Medicaid managed care/HealthChoices) mental health service. Includes both clinical services, such as individual and group therapy, and non-clinical services such as case management and peer support

Table A.6: Descriptions of adverse outcomes included in the aggregated indicator variable used for evaluation in Table 5.4. If any of these outcomes occurred for the child within 2 years of the current referral, we assign the aggregate indicator a positive value (positive for 1,762 out of 14,417 test data points).

	All (14,417)		Black (6,841)		White (5,763)	
	Struct.	Hybrid	Struct.	Hybrid	Struct.	Hybrid
AUC	75.75	76.25*	74.69	75.65*	74.93	74.97
TPR	56.09	56.32*	56.93	57.39*	53.87*	53.44
FPR	19.58	19.52*	20.82	20.39*	19.65*	19.98
Precision	32.49	32.66*	37.04	37.72*	28.20*	27.70
F1	41.14	41.34*	44.88	45.52*	37.02*	36.48
Accuracy	76.92	77.01*	75.24	75.68*	77.03*	76.69

Table A.7: Metrics for AFST task with different models, when incorporating previous referral notes, over all data. Where the difference between the hybrid and structured models is significant ($p < 0.05$) the better-performing value is starred.

	All (4,133)		Black (1,880)		White (1,894)	
	Struct.	Hybrid	Struct.	Hybrid	Struct.	Hybrid
AUC	69.79	71.83*	68.24	70.54*	70.03	71.90*
TPR	59.61	70.10*	59.72	71.48*	58.49	67.49*
FPR	31.84*	40.38	33.50*	43.59	30.72*	37.43
Precision	31.37*	29.77	36.48*	34.57	26.54*	25.49
F1	41.11	41.79*	45.29	46.59*	36.51	37.00*
Accuracy	66.48*	61.68	64.85*	60.08	67.56*	63.36

Table A.8: Metrics for AFST task with different models, when incorporating previous referral notes, over test data that contains text. Where the difference between the hybrid and structured models is significant ($p < 0.05$) the better-performing value is starred.

Bibliography

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260.

Julia Adams, Hannah Brückner, and Cambria Naslund. 2019. Who counts as a notable sociologist on Wikipedia? gender, race, and the "professor test". *Socius*, 5.

Saifuddin Ahmed, Kokil Jaidka, and Jaeho Cho. 2016. The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties. *Telematics and Informatics*, 33(4):1071–1087.

Beatrice Alex, Clare Llewellyn, Claire Grover, Jon Oberlander, and Richard Tobin. 2016. Homing in on Twitter users: Evaluating an enhanced geoparser for user profile locations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3936–3944, Portorož, Slovenia. European Language Resources Association (ELRA).

Michelle Alexander. 2011. The new jim crow. *Ohio St. J. Crim. L.*, 9:7.

Aerielle M Allen and Colin Wayne Leach. 2018. The psychology of Martin Luther King Jr.'s "creative maladjustment" at societal injustice and oppression. *Journal of Social Issues*, 74(2):317–336.

Silvio Amir, Jan-Willem van de Meent, and Byron Wallace. 2021. On the impact of random seeds on the fairness of clinical classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3808–3823, Online. Association for Computational Linguistics.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2019. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. 2016. *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*.

Anjuman Antil and Harsh V. Verma. 2019. Rahul Gandhi on Twitter: An analysis of brand building through Twitter by the leader of the main opposition party in India. *Global Business Review*, 0(0).

Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the part: Examining information operations within #blacklivesmatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, 2.

Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *Proc. of ASONAM*, pages 258–265.

Kimberly F. Balsam, Yamile Molina, Blair Beadnell, Jane Simoni, and Karina Walters. 2011. Measuring multiple minority stress: The lgbt people of color microaggressions scale. *Cultur Divers Ethnic Minor Psychol.*, 17(2):163–174.

David Bamman. 2015. *People-Centric Natural Language Processing.* Ph.D. thesis, Carnegie Mellon University.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

David Bamman and Noah A Smith. 2014. Unsupervised discovery of biographical structure from text. *TACL*, 2.

Antoine J Banks, Ismail K White, and Brian D McKenzie. 2019. Black politics: How anger influences the political actions blacks pursue to reduce racial inequality. *Political behavior*, 41(4):917–943.

Daniel Bar-Tal, Carl F Graumann, Arie W Kruglanski, and Wolfgang Stroebe. 2013. *Stereotyping and prejudice: Changing conceptions.* Springer Science & Business Media.

John Bargh. 1999. The cognitive monster: The case against the controllability of automatic stereotype effects. *Dual-process theories in social psychology*, pages 361–382.

Lisa Feldman Barrett and James A Russell. 2014. *The psychological construction of emotion.* Guilford Publications.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482.

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, 37(1):38–44.

Derrick A Bell Jr. 1980. Brown v. board of education and the interest-convergence dilemma. *Harvard law review*, pages 518–533.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623, New York, NY, USA. Association for Computing Machinery.

Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Michael Biggs. 2018. Size matters: Quantifying protest by counting participants. *Sociological Methods & Research*, 47(3):351–383.

Irene V. Blair. 2002. The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, 6(3):242–261.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR*, 3.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. 2021. An interpretability illusion for BERT. *arXiv preprint arXiv:2104.07143*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Yarimar Bonilla and Jonathan Rosa. 2015. #ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States. *American Ethnologist*, 42(1):4–17.

Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind egyptian political polarization on twitter. In *Proc. of CSCW*, page 700–711.

Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. Inside Technology. MIT Press.

Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2013. Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop, London*.

Samantha Bradshaw, Hannah Bailey, and P Howard. 2021. Industrialized disinformation: 2020 global inventory of organized social media manipulation. computational propaganda research project.

Samantha Bradshaw and Philip N Howard. 2018. Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Project*.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–12, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 2020 Conference on Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Simone Browne. 2015. *Dark Matters: On the Surveillance of Blackness*. Duke University Press.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4-5).

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91.

Children's Bureau. 2017. Making and screening reports of child abuse and neglect.

Camille D Burge. 2020. Introduction to dialogues: Black affective experiences in politics. *Politics, Groups, and Identities*, 8(2):390–395.

Radzhana Buyantueva. 2018. Lgbt rights activism and homophobia in russia. *Journal of Homosexuality*, 65(4):456–483. PMID: 28409697.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ewa S Callahan and Susan C Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10).

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Lindsay Cattell, Julie Bruch, et al. 2021. Identifying students at risk using prior performance versus a machine learning algorithm. Technical report, Mathematica Policy Research.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. In *Proc. of CSCW*.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 2009 Conference on Advances in Neural Information Processing Systems*, pages 288–296.

Erica Chenoweth and Maria J Stephan. 2011. *Why civil resistance works: The strategic logic of nonviolent conflict*. Columbia University Press.

Judeth Oden Choi, James Herbsleb, Jessica Hammer, and Jodi Forlizzi. 2020. Identity-based roles in rhizomatic social justice movements on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 488–498.

Munmun Choudhury, Shagun Jhaver, Benjamin Sugar, and Ingmar Weber. 2016. Social media participation in an activist movement for racial equality. *Proceedings of the ... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, 2016:92–101.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pages 134–148, New York, NY, USA. PMLR.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Heather M Clarke and Kara A Arnold. 2018. The influence of sexual orientation on the perceived fit of male applicants for both male- and female-typed jobs. *Frontiers in Psychology*, 9:656.

B.C. Cohen. 1963. *Press and Foreign Policy*. Princeton Legacy Library. Princeton University Press.

P.H. Collins. 1990. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Perspectives on Gender. Taylor & Francis.

Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 582–593, New York, NY, USA. Association for Computing Machinery.

Kate Crawford et al. 2021. Time to regulate ai that interprets human emotions. *Nature*, 592(7853):167–167.

Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. of Chicago Legal Forum*, 1989(8).

Crowd Counting Consortium. 2022. crowdcounting.org. Accessed January 24, 2022.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st International Conference on World Wide Web*, page 699–708, New York, NY, USA. Association for Computing Machinery.

Philipp Darius and Fabian Stephany. 2019. Twitter "Hashjacked": Online polarisation strategies of Germany's political far-right. In *Proceedings of the 2019 International Conference on Social Informatics*, pages 188–201.

Kareem Darwish. 2019. Quantifying polarization on twitter: The Kavanaugh nomination. In *Proceedings of the 2019 International Conference on Social Informatics*, pages 188–201.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515.

Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2021. Leveraging expert consistency to improve algorithmic decision support. *2021 Workshop on Information Technologies and Systems*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 120–128, New York, NY, USA. Association for Computing Machinery.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why cohen's kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Diane DePanfilis. 2003. *Child protective services: A guide for caseworkers*. US Department of Health and Human Services, Administration for Children.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Alan J Dettlaff, Stephanie L Rivaux, Donald J Baumann, John D Fluke, Joan R Rycraft, and Joyce James. 2011. Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare. *Children and Youth Services Review*, 33(9):1630–1637.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Dieterich, Christina Mendoza, and Tim Brennan. 2016. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, 7(7.4):1.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, page 67–73, New York, NY, USA. Association for Computing Machinery.

Kanae Doi and Philippa Stewart. 2019. Interview: The invisible struggle of Japan's transgender population. *Human Rights Watch*.

MeiXing Dong, David Jurgens, Carmen Banea, and Rada Mihalcea. 2019. Perceptions of social roles across cultures. In *Proceedings of the 2019 International Conference on Social Informatics*, pages 157–172.

Andrea Lane Eastman, Lisa Schelbe, and Jacquelyn McCroskey. 2019. A content analysis of case records: Two-generations of child protective services involvement. *Children and Youth Services Review*, 99:308–318.

Jennifer L Eberhardt. 2020. *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. Penguin Books.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *Working Paper*.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Young-Ho Eom, Pablo Aragón, David Laniado, Andreas Kaltenbrunner, Sebastiano Vigna, and Dima L Shepelyansky. 2015. Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. *PloS one*, 10(3).

Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the 2017 Conference of the Association for the Advancement of Artificial Intelligence*, pages 963–969.

Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, pages 112–120.

Paolo Ferragina, Francesco Piccinno, and Roberto Santoro. 2015. On analyzing hashtags in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 110–119.

Muhammad Feyyaz. 2019. Contextualizing the Pulwama attack in Kashmir–a perspective from Pakistan. *Perspectives on Terrorism*, 13(2):69–74.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: A case study of people portrayals in online #MeToo stories. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):158–169.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Anjalie Field, Amanda Coston, Alexandra Chouldechova, David Steier, and Yulia Tsvetkov. Forthcoming(a). The opportunities and pitfalls of using natural language processing for risk prediction: A case study in the child welfare system.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Anjalie Field, Chan Young Park, Kevin Z. Lin, and Yulia Tsvetkov. 2022. Controlled analyses of social biases in Wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, page 2624–2635, New York, NY, USA. Association for Computing Machinery.

Anjalie Field, Antonio Theophilo, Chan Young Park, Jamelle Watson-Daniels, and Yulia Tsvetkov. Forthcoming(b). Emotion analysis and the role of positivity in #BlackLivesMatter tweets.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2550–2560, Florence, Italy. Association for Computational Linguistics.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608, Online. Association for Computational Linguistics.

Charles J Fillmore. 1982. Frame semantics. *Cognitive linguistics: Basic readings*, pages 373–400.

Andrew R. Flores. 2019. Social acceptance of lgbt people in 174 countries. https://williamsinstitute.law.ucla.edu/publications/global-acceptance-index-lgbt/. Accessed: 2021-04-17.

Riccardo Fogliato, Alice Xiang, Zachary Lipton, Daniel Nagin, and Alexandra Chouldechova. 2021. On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 100–111, New York, NY, USA. Association for Computing Machinery.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4).

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, volume 12.

Deen Freelon, Charlton D McIlwain, and Meredith Clark. 2016. Beyond the hashtags: #Ferguson,#BlackLivesMatter, and the online struggle for offline justice. *Center for Media & Social Impact, American University*.

John R. French and Bertram Raven. 1959. The bases of social power. *Studies in Social Power*, page 150–167.

Batya Friedman, Peter Kahn, Alan Borning, and Alina Huldtgren. 2013. Value sensitive design and information systems. In Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael Gorman, editors, *Early engagement and new technologies: Opening up the laboratory*, volume 16. Springer, Dordrecht.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24, Florence, Italy. Association for Computational Linguistics.

Roland G Fryer Jr and Steven D Levitt. 2004. The causes and consequences of distinctively black names. *The Quarterly Journal of Economics*, 119(3):767–805.

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: using language models to quantify gender bias in sports journalism. *IJCAI workshop on NLP meets Journalism.*

William A Gamson and Andre Modigliani. 1989. Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*, 95(1):1–37.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Siddhant Garg, Rohit Kumar Sharma, and Yingyu Liang. 2020. Beyond fine-tuning: Few-sample sentence embedding transfer. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 460–469, Suzhou, China. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Salma I Ghanem and Maxwell McCombs. 2001. The convergence of agenda setting and framing. In *Framing public life*, pages 83–98. Routledge.

Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, page 1037–1046, New York, NY, USA. Association for Computing Machinery.

Andrew B Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Workshop on Graph Based Methods for Natural Language Processing*.

Claudia Goldin. 1990. *Understanding the gender gap: an economic history of American women*. NBER series on long-term factors in economic development. Oxford University Press.

Yevgeniy Golovchenko, Mareike Hartmann, and Rebecca Adler-Nissen. 2018. State, media and civil society in the information warfare over Ukraine: citizen curators of digital disinformation. *International Affairs*, 94(5):975–994.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems*, pages 2672–2680.

Jeff Goodwin, James M. Jasper, and Francesca Polletta. 2007. *Emotional Dimensions of Social Movements*, chapter 18. John Wiley & Sons, Ltd.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in Wikipedia. In *Proc. of Hypertext & Social Media*.

Clive WJ Granger. 1988. Some recent development in a concept of causality. *Journal of econometrics*, 39(1-2):199–211.

Ben Green. 2019. "good" isn't good enough. In *Proceedings of the AI for Social Good Workshop*, pages 7–14.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Xing Sam Gu and Paul R. Rosenbaum. 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

Itika Gupta, Barbara Di Eugenio, Devika Salunke, Andrew Boyd, Paula Allen-Meares, Carolyn Dickens, and Olga Garcia. 2020. Heart failure education of African American and Hispanic/Latino patients: Data collection and analysis. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 41–46, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

David L Hamilton and Tina K Trolier. 1986. Stereotypes and stereotyping: An overview of the cognitive approach in prejudice, discrimination, and racism. *Prejudice, discrimination, and racism.*

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Valerie S Harder, Elizabeth A Stuart, and James C Anthony. 2010. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, 15(3).

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331.

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. 2017. Demographic word embeddings for racism detection on Twitter. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 926–936, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1803–1812, New York, NY, USA. Association for Computing Machinery.

US Department of Health and Human Services. 2017. Child maltreatment 2017. Technical report, Children's Bureau (Ed.).

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 237–246, New York, NY, USA. Association for Computing Machinery.

David R Heise. 1979. *Understanding events: Affect and the construction of social action*. CUP Archive.

David R Heise. 2007. *Expressive order: Confirming sentiments in social actions*. Springer Science & Business Media.

Aurélie Herbelot, Eva von Redecker, and Johanna Müller. 2012. Distributional techniques for philosophical enquiry. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 45–54, Avignon, France. Association for Computational Linguistics.

Robert B Hill. 2004. Institutional racism in child welfare. *Race and Society*, 7(1):17–33.

Robert B Hill. 2005. The role of race in foster care placements. *Race matters in child welfare: The overrepresentation of African American children in the system*, pages 187–200.

Laura Hollink, Astrid van Aggelen, and Jacco van Ossenbruggen. 2018. Using the web of data to study gender differences in online knowledge sources: The case of the european parliament. In *Proceedings of the 10th ACM Conference on Web Science*, page 381–385, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Chao-Chun Hsu, Shantanu Karnwal, Sendhil Mullainathan, Ziad Obermeyer, and Chenhao Tan. 2020. Characterizing the value of information in medical notes. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2062–2072, Online. Association for Computational Linguistics.

Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, page 173–178, New York, NY, USA. Association for Computing Machinery.

Thomas Huckin. 2002. Critical discourse analysis and the discourse of condescension. *Discourse studies in composition*, pages 155–176.

Leonie Huddy, Stanley Feldman, and Christopher Weber. 2007. The political consequences of perceived threat and felt insecurity. *The ANNALS of the American Academy of Political and Social Science*, 614(1):131–153.

Anne H. Charity Hudley. 2017. Language and racialization. In Ofelia García, Nelson Flores, and Massimiliano Spotti, editors, *The Oxford Handbook of Language and Society*, pages 381–402. Oxford University Press.

Anne H Charity Hudley, Christine Mallinson, and Mary Bucholtz. 2020. Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language*, 96(4):e200–e235.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.

Sarah J. Jackson, Moya Bailey, and Brooke Foucault Welles. 2020. *#HashtagActivism: Networks of Race and Gender Justice*. The MIT Press.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, page 375–385. Association for Computing Machinery.

James M. Jasper. 1997. *The Art of Moral Protest: Culture, Biography, and Creativity in Social Movements*. University of Chicago Press.

James M. Jasper. 2011. Emotions and social movements: Twenty years of theory and research. *Annual Review of Sociology*, 37(1):285–303.

Shaoxiong Ji, Shirui Pan, and Pekka Marttinen. 2021. Medical code assignment with gated convolution and note-code interaction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1034–1043, Online. Association for Computational Linguistics.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada. Association for Computational Linguistics.

May Jiang and Christiane Fellbaum. 2020. Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25, Barcelona, Spain (Online). Association for Computational Linguistics.

Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. 2020. Factoring fact-checks: Structured information extraction from fact-checking articles. In *Proceedings of The Web Conference 2020*, page 1592–1603, New York, NY, USA. Association for Computing Machinery.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Adam Jones. 2002. The russian press in the post-soviet era: a case study of izvestia. *Journalism Studies*, 3(3):359–375.

Kenneth Joseph and Jonathan Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online. Association for Computational Linguistics.

Kenneth Joseph, Wei Wei, and Kathleen M. Carley. 2017. Girls rule, boys drool: Extracting semantic and affective stereotypes from Twitter. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, page 1362–1374, New York, NY, USA. Association for Computing Machinery.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dan Jurafsky, Victor Chahuneau, Bryan Routledge, and Noah Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4).

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't use English dev: On the zero-shot cross-lingual evaluation of contextual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 549–554, Online. Association for Computational Linguistics.

Gary King, Jennifer Pan, and Margaret E Roberts. 2013. How censorship in china allows government criticism but silences collective expression. *American political science Review*, 107(2):326–343.

Gary King, Jennifer Pan, and Margaret E Roberts. 2017. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, 111(3):484–501.

Sharese King. 2020. From African American Vernacular English to African American Language: Rethinking the study of race and language in African Americans' speech. *Annual Review of Linguistics*, 6(1):285–300.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Kluttz, Joshua A Kroll, Jenna Burrell, and Deirdre Mulligan. 2018. Afog workshop panel 1: What a technical 'fix'for fairness can and can't accomplish. Technical report, Algorithmic Fairness & Opacity Working Group.

Nancy Krieger. 1990. Racial and gender discrimination: risk factors for high blood pressure? *Social science & medicine*, 30(12):1273–1281.

Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory approaches to machine learning. International Conference on Machine Learning Workshop.

Amit Kumar, Somesh Dhamija, and Aruna Dhamija. 2016. Political marketing: The horizon of present era politics. *SCMS Journal of Indian Management*, 13(4):116–125.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4153–4163, Hong Kong, China. Association for Computational Linguistics.

P Kumaraguru, S Singh, D Manu, K Gupta, A Sadaria, S Srikanth, H Bhatia, S Garimella, K. Buggana, A Agarwal, A Kapoor, K Gupta, T Garg, O Gurjar, and S Saini. 2019. Social media to win elections: Analysis of #LokSabhaElections2019 in India. *Precog Technical report.*

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Virgile Landeiro, Tuan Tran, and Aron Culotta. 2019. Discovering and controlling for latent confounds in text classification using adversarial domain adaptation. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 298–305. SIAM.

Wendy G Lane, David M Rubin, Ragin Monteith, and Cindy W Christian. 2002. Racial differences in the evaluation of pediatric fractures for physical abuse. *Jama*, 288(13):1603–1609.

Isabelle Langrock and Sandra González-Bailón. 2020. The gender divide in Wikipedia: A computational approach to assessing the impact of two feminist interventions. *Available at SSRN*.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Computational social science. *Science*, 323(5915):721–723.

Huyen Le, GR Boynton, Zubair Shafiq, and Padmini Srinivasan. 2019. A postmortem of suspended Twitter accounts in the 2016 US presidential election. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pages 258–265.

Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Joy Leopold and Myrtle P Bell. 2017. News media and the racialization of protest: An analysis of Black Lives Matter articles. *Equality, Diversity and Inclusion: An International Journal*.

Michael Lepori. 2020. Unequal representations: Analyzing intersectional biases in word embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1720–1728, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Alexa Lisitza. 2017. History of pride parades in the u.s. *TeenVogue*.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 5.

Kirsten Lloyd. 2018. Bias amplification in artificial intelligence systems. *CoRR*, abs/1809.07842.

Christine Logel, Emma C. Iserman, Paul G. Davies, Diane M. Quinn, and Steven J. Spencer. 2009. The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology*, 45(2):299 – 312.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

Miranda L. Y. Ma. 2017. Affective framing and dramaturgical actions in social movements. *Journal of Communication Inquiry*, 41(1):5–21.

Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016. #ISISisNotIslam or #DeportAllMuslims? predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*, page 95–106, New York, NY, USA. Association for Computing Machinery.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, pages 173–182. ACM.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Maxwell McCombs. 2002. The agenda-setting role of the mass media in the shaping of public opinion. In *Proceedings of the 2002 Conference of Mass Media Economics, London School of Economics.*

Duncan McDonnell and Luis Cabrera. 2019. The right-wing populism of India's Bharatiya Janata Party (and why comparativists should care). *Democratization*, 26(3):484–501.

Charlton D. McIlwain. 2019. *Black Software: The Internet and Racial Justice, from the AfroNet to Black Lives Matter*. Oxford University Press, Incorporated.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29).

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3:55.

Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. *Computing Research Repository*, arXiv:1901.10436. Version 6.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of American football broadcasts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6355–6361, Hong Kong, China. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *In Proceedings of the 2013 International Conference on Learning Representations.*

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

Satish Mishra. 2019. Emerging electoral dynamics after Pulwama tragedy. *Observer Research Foundation.*

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, page 85–94, New York, NY, USA. Association for Computing Machinery.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Marçal Mora-Cantallops, Salvador Sánchez-Alonso, and Elena García-Barriocanal. 2019. A systematic literature review on wikidata. *Data Technologies and Applications*, 53(3).

Suhanthie Motha. 2020. Is an antiracist and decolonizing applied linguistics possible? *Annual Review of Applied Linguistics*, 40:128–133.

Kevin Munger, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2018. Elites tweet to get feet off the streets: Measuring regime social media strategies during protest. *Political Science Research and Methods*, pages 1–20.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Michael Nash. 2017. Examination of using structured decision making and predictive analytics in assessing safety and risk in child welfare. *Los Angeles: County of Los Angeles Office of Child Protection*.

National Center for Science and Engineering Statistics. 2019. Doctorate recipients from U.S. universities. National Science Foundation.

Viet-An Nguyen, Jordan L Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of the 2013 Conference on Advances in Neural Information Processing Systems*, pages 1106–1114.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Safiya Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Rodrigo Ochigame and Katherine Ye. 2021. Search atlas: Visualizing divergent search results across geopolitical borders. In *Designing Interactive Systems Conference 2021*, page 1970–1983, New York, NY, USA. Association for Computing Machinery.

Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for hci. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–16, New York, NY, USA. Association for Computing Machinery.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Alexandra Olteanu, Ingmar Weber, and Daniel Gatica-Perez. 2015. Characterizing the demographics behind the #blacklivesmatter movement. In *Proceedings of AAAI Spring Symposia on Observational Studies through Social Media and Other Human-Generated Content*, pages 4144–4154.

Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

C.E. Osgood, G.J. Suci, and P.H. Tannenbaum. 1957. *The Measurement of Meaning*. Illini Books, IB47. University of Illinois Press.

S. O'Connell. 2016. DPHHS monitoring: Child and family services division state laws on emergency removal of children. *Montana Children, Families, Health, and Human Services Interim Committee*.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2020. Hope speech detection: A computational analysis of the voice of peace. In *Proceedings of the 24th European Conference on Artificial Intelligence*.

Abhinav Pandya. 2019. The future of Indo-Pak relations after the Pulwama attack. *Perspectives on Terrorism*, 13(2):65–68.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Raghavendra Reddy Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. 2021. Multilingual contextual affective analysis of LGBT people portrayals in Wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 479–490.

Amandalynne Paullada. 2020. How Does Machine Translation Shift Power? In *Proceedings of the First Workshop on Resistance AI*.

Wendy Pearlman. 2013. Emotions and the microfoundations of the arab uprisings. *Perspectives on Politics*, 11(2):387–409.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*.

Brian E Perron, Bryan G Victor, Gregory Bushman, Andrew Moore, Joseph P Ryan, Alex Jiahong Lu, and Emily K Piellusch. 2019. Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. *Child abuse & neglect*, 98:104180.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Davin L Phoenix. 2019. *The anger gap: How race shapes emotion in politics*. Cambridge University Press.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online. Association for Computational Linguistics.

Vinodkumar Prabhakaran and Owen Rambow. 2017. Dialog structure through the lens of gender, gender environment, and power. *Dialogue & Discourse*, 8(2):21–55.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1534–1545, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Reid Pryzant, Youngjoo Chung, and Dan Jurafsky. 2017. Predicting sales from the language of product descriptions. In *Proceedings of the Special Interest Group on Information Retrieval (SIGIR) eCommerce Workshop*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:480–489.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report, OpenAI*.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660.

Yolanda A. Rankin and Jakita O. Thomas. 2019. Straighten up and fly right: Rethinking intersectionality in hci research. *Interactions*, 26(6):64–68.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Mohammad Sadegh Rasooli, Noura Farra, Axinia Radeva, Tao Yu, and Kathleen McKeown. 2018. Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1):143–165.

Joseph Reagle and Lauren Rhue. 2011. Gender bias in Wikipedia and Britannica. *International Journal of Communication*, 5.

Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2020. A taxonomy of knowledge gaps for Wikimedia projects (second draft). *arXiv preprint arXiv:2008.12314*.

Eugenia Ha Rim Rho, Gloria Mark, and Melissa Mazmanian. 2018. Fostering civil discourse online: Linguistic behavior in comments of #metoo articles across political perspectives. *Proceedings of the ACM Conference on Human-Computer Interaction*, 2(CSCW).

Filipe N Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth international AAAI conference on web and social media*.

Allissa V Richardson. 2019. Dismantling respectability: The rise of new womanist communication models in the era of black lives matter. *Journal of Communication*, 69(2):193–213.

Rashida Richardson, Jason M Schultz, and Kate Crawford. 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online*, 94:15.

Dorothy Roberts. 2009. *Shattered bonds: The color of child welfare*. Civitas Books.

Dorothy E Roberts. 2019. Digitizing the carceral state. *Harvard Law Review*, 132.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science (Forthcoming)*, 64.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, and Edoardo M Airoldi. 2013. The structural topic model and applied social science. In *Presented at the NeurIPS Workshop on Topic Models: Computation, Application, and Evaluation*, pages 1–20.

Dawn T Robinson, Lynn Smith-Lovin, and Allison K Wisecup. 2006. Affect control theory. In *Handbook of the sociology of emotions*, pages 179–202. Springer.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46(5):621–647.

Paul R. Rosenbaum. 1988. Sensitivity analysis for matching with multiple controls. *Biometrika*, 75(3):577–581.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1).

Paul R. Rosenbaum and Donald B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39.

Wendy D Roth. 2016. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8).

Arturas Rozenas and Denis Stukal. 2019. How autocrats manipulate economic news: Evidence from russia's state-controlled television. *The Journal of Politics*, 81(3):982–996.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Jessica M. Salerno and Liana C. Peter-Hagene. 2013. The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science*, 24(10):2069–2078. PMID: 23969778.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

P.K. Saucier, T.P. Woods, P. Douglass, B. Hesse, T.K. Nopper, G. Thomas, and C. Wun. 2016. *Conceptual Aphasia in Black: Displacing Racial Formation*. Critical Africana Studies. Lexington Books.

Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the us child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional hci: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 5412–5427, New York, NY, USA. Association for Computing Machinery.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Jamil S Scott and Jonathan Collins. 2020. Riled up about running for office: examining the impact of emotions on political ambition. *Politics, Groups, and Identities*, 8(2):407–422.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Anuj K Shah and Daniel M Oppenheimer. 2008. Heuristics made easy: an effort-reduction framework. *Psychological bulletin*, 134(2):207.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the Tenth international AAAI conference on web and social media*.

Prabhsimran Singh, Kuldeep Kumar, Karanjeet Singh Kahlon, and Ravinder Singh Sawhney. 2019. Can tweets predict election results? insights from twitter analytics. In *Advanced Informatics for Computing Research*, pages 271–281, Singapore. Springer Singapore.

Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 21–29, New York, NY, USA. Association for Computing Machinery.

M Sloane, E Moss, O Awomolo, and L Forlano. 2020. Participation is not a design fix for machine learning. *Computing Research Repository*, arXiv:2007.02423. Version 3.

Anton Sobolev, M. Keith Chen, Jungseock Joo, and Zachary C. Steinert-Threlkeld. 2020. News and geolocated social media accurately measure protest size variation. *American Political Science Review*, 114(4):1343–1351.

Pia Sommerauer and Antske Fokkens. 2019. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233, Florence, Italy. Association for Computational Linguistics.

Tami Spry. 1995. In the absence of word and body: Hegemonic implications of "victim" and "survivor" in women's narratives of sexual violence. *Women and Language*, 13(2):27.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM Conference on Human-Computer Interaction*, 3:1–26.

Claude M Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology*, 69(5):797.

Zachary Steinert-Threlkeld and Jungseock Joo. 2020. Protest event data from geolocated social media content.

Leo Graiden Stewart, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird. 2017. Drawing the lines of contention: Networked frame contests within #blacklivesmatter discourse. *Proceedings of ACM Conference on Human-Computer Interactation*, 1.

Fritz Strack and Thomas Mussweiler. 1997. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, 73(3):437.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Elizabeth Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Stat Sci*, 25(1).

Derald Wing Sue. 2010. *Microaggressions in everyday life: Race, gender, and sexual orientation.* John Wiley & Sons.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue*, 11(3):10–29.

Yi Chern Tan and L Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Proceedings of the 2019 Conference on Advances in Neural Information Processing Systems*, volume 32, pages 13230–13241. Curran Associates, Inc.

Rachael Tatman. 2020. What I Won't Build. Workshop on Widening NLP.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

András Tilcsik, Michel Anteby, and Carly R. Knight. 2015. Concealable stigma and occupational segregation: Toward a theory of gay and lesbian occupations. *Administrative Science Quarterly*, 60(3):446–481.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1629–1638.

Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M Carley. 2020. A computational analysis of polarization on Indian and Pakistani social media. In *Proceedings of the 2020 International Conference on Social Informatics*, pages 364–379. Springer.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation. *Center for Social data Analytics*.

Nicholas A Valentino, Ted Brader, Eric W Groenendyk, Krysha Gregorowicz, and Vincent L Hutchings. 2011. Election night's alright for fighting: The role of emotions in political participation. *The Journal of Politics*, 73(1):156–170.

Nicholas A Valentino, Krysha Gregorowicz, and Eric W Groenendyk. 2009. Efficacy, emotions and the habit of participation. *Political Behavior*, 31(3):307–330.

Marcel H Van Herpen. 2015. *Putin's propaganda machine: Soft power and russian foreign policy*. Rowman & Littlefield.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR.

Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, page 941–953, New York, NY, USA. Association for Computing Machinery.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10).

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1).

J. Celeste Walley-Jean. 2009. Debunking the myth of the "angry black woman": An exploration of anger in young african american women. *Black Women, Gender + Families*, 3(2):68–86.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019*.

Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant document discovery for fact-checking articles. In *Companion Proceedings of the The Web Conference 2018*, page 525–533, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zijian Wang and Christopher Potts. 2019. TalkDown: A corpus for condescension detection in context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3711–3719, Hong Kong, China. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Zeerak Waseem, Smarika Lulz, and Isabelle Bingel, Joachim Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. *Computing Research Repository*, arXiv:2101.11974. Version 1.

Ingmar Weber, Venkata R. Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. Islamist polarization in Egypt on Twitter. In *Proc. of ASONAM*, pages 290—297.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford. 2019. Gendered ambiguous pronoun (GAP) shared task at the gender bias in NLP workshop 2019. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Florence, Italy. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Susan J Wells, Lani M Merritt, and Harold E Briggs. 2009. Bias, racism and evidence-based practice: The case for more focused development of the child welfare evidence base. *Children and Youth Services Review*, 31(11):1160–1171.

Funk MJ Westreich D, Lessler J. 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*, 63(8):826–833.

Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI now report 2018*. AI Now Institute at New York University New York.

Cai Wilkinson. 2014. Putting “traditional values” into practice: The rise and contestation of anti-homopropaganda laws in russia. *Journal of Human Rights*, 13(3):363–379.

Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing twitter data in social research: Taking into account users’ views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168. PMID: 29276313.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Donghyeon Won, Zachary C. Steinert-Threlkeld, and Jungseock Joo. 2017. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 786–794, New York, NY, USA. Association for Computing Machinery.

Zach Wood-Doughty, Nicholas Andrews, Rebecca Marvin, and Mark Dredze. 2018. Predicting Twitter user demographics from names alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 105–111, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 247–257, Tokyo, Japan. Association for Computational Linguistics.

Amber Young, Ari D Wigdor, and Gerald Kane. 2016. It's not what you think: Gender bias in information about fortune 1000 CEOs on Wikipedia. In *Proceedings of the International Conference for Information Systems*.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: Quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, page 110–120, New York, NY, USA. Association for Computing Machinery.

Zhongheng Zhang, Hwa Jung Kim, Guillaume Lonjon, and Yibing Zhu. 2019. Balance diagnostics after propensity score matching. *Annals of translational medicine*, 7.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.