Open Extraction of Fine-Grained Political Statements

David Bamman
School of Information
University of California, Berkeley
Berkeley, CA 94720, USA
dbamman@berkeley.edu

Noah A. Smith
Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
nasmith@cs.washington.edu

Abstract

Text data has recently been used as evidence in estimating the political ideologies of individuals, including political elites and social media users. While inferences about people are often the intrinsic quantity of interest, we draw inspiration from open information extraction to identify a new task: inferring the political import of propositions like OBAMA IS A SOCIALIST. We present several models that exploit the structure that exists between people and the assertions they make to learn latent positions of people and propositions at the same time, and we evaluate them on a novel dataset of propositions judged on a political spectrum.

1 Introduction

Over the past few years, much work has focussed on inferring political preferences of people from their behavior, both in unsupervised and supervised settings. Classical ideal point models (Poole and Rosenthal, 1985; Martin and Quinn, 2002) estimate the political ideologies of legislators through their observed voting behavior, possibly paired with the textual content of bills (Gerrish and Blei, 2012) and debate text (Nguyen et al., 2015); other unsupervised models estimate ideologies of politicians from their speeches alone (Sim et al., 2013). Twitter users have also been modeled in a similar framework, using their observed following behavior of political elites as evidence to be explained (Barberá, 2015). Supervised models, likewise, have not only been used for assessing the political stance of sentences (Iyyer et al., 2014) but are also very popular for predicting the holistic ideologies of everyday users on Twitter (Rao et al., 2010; Pennacchiotti and Popescu, 2011; Al Zamal et al., 2012; Cohen and Ruths, 2013; Volkova et al., 2014), Facebook (Bond and Messing, 2015) and blogs (Jiang and Argamon, 2008), where training data is relatively easy to obtain—either from user self-declarations, political following behavior, or third-party categorizations.

Aside from their intrinsic value, estimates of users' political ideologies have been useful for quantifying the orientation of news media sources (Park et al., 2011; Zhou et al., 2011). We consider in this work a different task: estimating the political import of propositions like OBAMA IS A SOCIALIST.

In focusing on propositional statements, we draw on a parallel, but largely independent, strand of research in open information extraction. IE systems, from early slot-filling models with predetermined ontologies (Hobbs et al., 1993) to the large-scale open-vocabulary systems in use today (Fader et al., 2011; Mitchell et al., 2015) have worked toward learning type-level propositional information from text, such as BARACK OBAMA IS PRESIDENT. To a large extent, the ability to learn these facts from text is dependent on having data sources that are either relatively factual in their presentation (e.g., news articles and Wikipedia) or are sufficiently diverse to average over conflicting opinions (e.g., broad, random samples of the web).

Many of the propositional statements that individuals make online are, of course, not objective descriptions of reality at all, but rather reflect their own beliefs, opinions and other private mental states (Wiebe et al., 2005). While much work has investigated methods for establishing the truth content of individual sentences — whether from the perspective of veridicality (de Marneffe et al., 2012), fact assessment (Nakashole and Mitchell, 2014), or subjectivity analysis (Wiebe et al., 2003; Wilson, 2008) — the structure that exists between users and their assertions gives us an opportunity to situate them both in the same political space: in this work we operate at the level of subject-

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 76–85,
Lisbon, Portugal, 17-21 September 2015.

predicate propositions, and present models that capture not only the variation in what subjects (e.g., OBAMA, ABORTION, GUN CONTROL) that individual communities are more likely to discuss, but also the variation in what predicates different communities assert of the same subject (e.g., GLOBAL WARMING IS A HOAX vs. IS A FACT). The contributions of this work are as follows:

- We present a new evaluation dataset of 766 propositions judged according to their positions in a political spectrum.
- We present and evaluate several models for estimating the ideal points of subject-predicate propositions, and find that unsupervised methods perform best (on sufficiently partisan data).

2 Task and Data

The task that we propose in this work is assessing the political import of type-level propositions; on average, are liberals or conservatives more likely to claim that GLOBAL WARMING IS A HOAX? To support this task, we create a benchmark of political propositions, extracted from politically partisan data, paired with human judgments (details in §2.3). We define a **proposition** to be a tuple comprised of a subject and predicate, each consisting of one or more words, such as ⟨global warming, *is a hoax*⟩.[1] We adopt an open vocabulary approach where each unique predicate defines a unary relation.

2.1 Data

In order to extract propositions that are likely to be political in nature and exhibit variability according to ideology, we collect data from a politically volatile source: comments on partisan blogs.

We draw data from NPR,[2] Mother Jones[3] and Politico[4], all listed by Pew Research (Mitchell et al., 2014) as news sources most trusted by those with consistently liberal views; Breitbart,[5] most trusted by those with consistently conservative views; and the Daily Caller,[6] Young Conservatives[7] and the Independent Journal Review,[8] all popular among conservatives (Kaufman, 2014). All data comes from articles published between 2012–2015 and is centered on the US political landscape.

Source	Articles	Posts	Tokens	Users
Politico	10,305	9.8M	348.4M	173,519
Breitbart	46,068	8.8M	336.4M	165,607
Daily Caller	46,114	5.4M	240.4M	228,696
Mother Jones	16,830	1.9M	119.2M	138,995
NPR	14993	1.6M	82.6M	62,600
IJ Review	3,396	278K	13.1M	51,589
Young Cons.	4,948	222K	10.6M	34,434
Total	142,654	28.0M	1.15B	621,231

Table 1: Data.

We gather comments using the Disqus API;[9] as a comment hosting service, Disqus allows users to post to different blogs using a single identity. Table 1 lists the total number of articles, user comments, unique users and tokens extracted from each blog source. In total, we extract 28 million comments (1.2 billion tokens) posted by 621,231 unique users.[10]

2.2 Extracting Propositions

The blog comments in table 1 provide raw data from which to mine propositional assertions. In order to extract structured ⟨subject, *predicate*⟩ propositions from text, we first parse all comments using the collapsed dependencies (de Marneffe and Manning, 2008) of the Stanford parser (Manning et al., 2014), and identify all subjects as those that hold an `nsubj` or `nsubjpass` relation to their head. In order to balance the tradeoff between generality and specificity in the representation of assertions, we extract three representations of each predicate.

1. Exact strings, which capture verbatim the specific nuance of the assertion. This includes all subjects paired with their heads and all descendants of that head. Tense and number are preserved.

   Example: ⟨Reagan, *gave amnesty to 3 million undocumented immigrants*⟩

2. Reduced syntactic tuples, which provide a level of abstraction by lemmatizing word forms and including only specific syntactic relationships. This includes propositions de-

[1] We use these typographical conventions throughout: Subjects are in sans serif, predicates in *italics*.

[2] `http://www.npr.org`

[3] `http://www.motherjones.com`

[4] `http://www.politico.com`

[5] `http://www.breitbart.com`

[6] `http://dailycaller.com`

[7] `http://www.youngcons.com`

[8] `https://www.ijreview.com`

[9] `https://disqus.com/api/`

[10] While terms of service prohibit our release of this data, we will make available tools to allow others to collect similar data from Disqus for these blogs.

fined as nominal subjects paired with their heads and children of that head that are negators, modal auxiliaries (*can, may, might, shall, could, would*), particles and direct objects. All word forms are lemmatized, removing tense information on verbs and number on nouns.

Example: ⟨Reagan, *give amnesty*⟩

3. Subject-verb tuples, which provide a more general layer of abstraction by only encoding the relationship between a subject and its main action. In this case, a proposition is defined as the nominal subject and its lemmatized head.

Example: ⟨Reagan, *give*⟩

The human benchmark defined in §2.3 below considers only verbatim predicates, while all models proposed in §3 and all baselines in §4 include the union of all three representations as data.

Here, syntactic structure not only provides information in the representation of propositions, but also allows us to define criteria by which to exclude predicates — since we are looking to extract propositions that are directly asserted by an author of a blog comment (and not second-order reporting), we exclude all propositions dominated by an attitude predicate (*Republicans think that Obama should be impeached*) and all those contained within a conditional clause (*If Obama were impeached...*). We also exclude all assertions drawn from questions (i.e., sentences containing a question mark) and all assertions extracted from quoted text (i.e., surrounded by quotation marks).

In total, from all 28 million comments across all seven blogs, we extract all propositions defined by the criteria above, yielding a total of 61 million propositions (45 million unique).

2.3 Human Benchmark

From all propositions with a verbatim predicate extracted from the entire dataset, we rank the most frequent subjects and manually filter out non-content terms (like *that*, *one*, *someone*, *anyone*, etc.) to yield a set of 138 target topics, the most frequent of which are *obama*, *democrats*, *bush*, *hillary*, and *america*.

For each proposition containing one of these topics as its subject and mentioned by at least 5 different people across all blogs, we randomly sampled 1,000 in proportion to their frequency of use (so that sentences that appear more frequently in the data are more likely to be sampled); the sentences selected in this random way contain a variety of politically charged viewpoints. We then presented them to workers on Amazon Mechanical Turk for judgments on the extent to which they reflect a US liberal vs. conservative political worldview.

For each sentence, we paid 7 annotators in the United States to a.) confirm that the extracted sentence was a well-formed assertion and b.) to rate "the most likely political belief of the person who would say it" on a five-point scale: very conservative/Republican (-2), slightly conservative/Republican (-1), neutral (0), slightly liberal/Democrat (1), and very liberal/Democrat (2).

We keep all sentences that at least six annotators have marked as meaningful (those excluded by this criterion include sentence fragments like *bush wasn't* and those that are difficult to understand without context, such as *romney is obama*) and where the standard deviation of the responses is under 1 (which excludes sentences with flat distributions such as *government does nothing well* and those with bimodal distributions, such as *christie is done*). After this quality control, we average the responses to create a dataset of 766 propositions paired with their political judgments. Table 2 presents a random sample of annotations from this dataset.

proposition	mean	s.d.
obama lied and people died	-2.000	0.000
gay marriage is not a civil right	-1.857	0.350
obama can't be trusted	-1.714	0.452
hillary lied	-0.857	0.990
hillary won't run	-0.714	0.452
bush was just as bad	0.857	0.639
obama would win	1.429	0.495
rand paul is a phony	1.429	0.495
abortion is not murder	1.571	0.495
hillary will win in 2016	1.857	0.350

Table 2: Random sample of AMT annotations.

3 Models

The models we introduce to assess the political import of propositions are based on two fundamental ideas. First, users' latent political preferences, while unobserved, can provide an organizing principle for inference about propositions in an unsupervised setting. Second, by decoupling the variation in *subjects* discussed by different communities (e.g., liberals may talk more

about global warming while conservatives may talk more about gun rights) from variation in what statements are *predicated* of those subjects (e.g., liberals may assert that ⟨global warming, *is a fact*⟩ while conservatives may be more likely to assert that it *is a hoax*), we are able to have a more flexible and interpretable parameterization of observed textual behavior that allows us to directly measure both.

We present two models below: one that represents users and propositions as real-valued points, and another that represents each as categorical variables. For both models, the input is a set of users paired with a list of ⟨subject, *predicate*⟩ tuples they author; the variables of interest we seek are representations of those users, subjects, and predicates that explain the coupling between users and propositions we see.

3.1 Additive Model

The first model we present (fig. 1) represents each user, subject, and predicate as a real-valued point in K-dimensional space. In the experiments that follow, we consider the simple case where $K = 1$ but present the model in more general terms below.

In this model, we parameterize the generative probability of a subject (like Obama) as used by an individual u as the exponentiated sum of a background log frequency of that subject in the corpus overall (m_{sbj}) and K additive effects, normalized over the space of S possible subjects, as a real-valued analogue to the SAGE model of Eisenstein et al. (2011). While the background term controls the overall frequency of a subject in the corpus, $\beta \in \mathbb{R}^{K \times S}$ mediates the relative increase or decrease in probability of a subject for each latent dimension. Intuitively, when both $\eta_{u,k}$ and $\beta_{k,sbj}$ (for a given user u, dimension k, and subject sbj) are the same sign (either both positive or both negative), the probability of that subject under that user increases; when they differ, it decreases. $\beta_{\cdot,sbj}$ is a K-dimensional representation of subject sbj, and $\eta_{u,\cdot}$ is a K-dimensional representation of user u.

$$P(sbj \mid u, \eta, \beta, m_{sbj}) = \frac{\exp\left(m_{sbj} + \sum_{k=1}^{K} \eta_{u,k}\beta_{k,sbj}\right)}{\sum_{sbj'} \exp\left(m_{sbj'} + \sum_{k=1}^{K} \eta_{u,k}\beta_{k,sbj'}\right)} \quad (1)$$

Likewise, we parameterize the generative probability of a predicate (conditioned on a subject) in the same way; for S subjects, each of which contains (up to) P predicates, $\psi \in \mathbb{R}^{S \times K \times P}$ captures the relative increase or decrease in probability for a given predicate conditioned on its subject, relative to its background frequency in the corpus overall, $m_{pred|sbj}$.

$$P(pred \mid sbj, u, \eta, \psi, m_{pred|sbj}) = \frac{\exp\left(m_{pred|sbj} + \sum_{k=1}^{K} \eta_{u,k}\psi_{sbj,k,pred}\right)}{\sum_{pred'} \exp\left(m_{pred'|sbj} + \sum_{k=1}^{K} \eta_{u,k}\psi_{sbj,k,pred'}\right)} \quad (2)$$

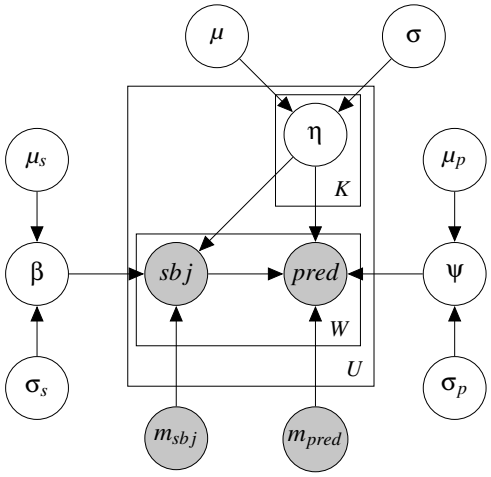

Figure 1: Additive model with decoupled subjects and predicates. η contains a K-dimensional representation of each user; β captures the variation in observed subjects, and ψ captures the variation in predicates for a fixed subject.

The full generative story for this model runs as follows. For a vocabulary of subjects of size S, where each subject s has P predicates:

- For each dimension k, draw subject coefficients $\beta_k \in \mathbb{R}^S \sim \text{Norm}(\mu_s, \sigma_s \mathbf{I})$
- For each subject s:
  - For each dimension k, draw subject-specific predicate coefficients $\psi_{s,k} \in \mathbb{R}^P \sim \text{Norm}(\mu_p, \sigma_p \mathbf{I})$
- For each user u:
  - Draw user representation $\eta \in \mathbb{R}^K \sim \text{Norm}(\mu, \sigma \mathbf{I})$
  - For each proposition $\langle sbj, pred \rangle$ made by u:
    - Draw sbj according to eq. 1
    - Draw $pred$ according to eq. 2

The unobserved quantities of interest in this model are η, β and ψ. In the experiments reported

below, we set the prior distributions on η, β and ψ to be standard normals ($\mu = 0, \sigma = 1$) and perform maximum *a posteriori* inference with respect to η, β and ψ in turn for a total of 25 iterations.

While β and ψ provide scores for the political import of subjects and of predicates conditioned on fixed subjects, respectively, we can recover a single ideological score for both a subject and its predicate by adding their effects together. In the evaluation given in §5, let the PREDICATE SCORE for $\langle$subject, *predicate*$\rangle$ be that given by $\psi_{\text{subject},\cdot,\textit{predicate}}$, and let the PROPOSITION SCORE be $\beta_{\cdot,\text{subject}} + \psi_{\text{subject},\cdot,\textit{predicate}}$.

3.2 Single Membership Model

While the additive model above represents each user and proposition as a real-valued point in K-dimensional space, we can also represent those values as categorical variables in an unsupervised naïve Bayes parameterization; in this case, a user is not defined as a mixture of different effects, but rather belongs to a single unique community. The generative story for this model (shown in fig. 2) is as follows:

- Draw population distribution over categories $\theta \sim \text{Dir}(\alpha)$
- For each category k, draw distribution over subjects $\phi_k \sim \text{Dir}(\gamma)$
- For each category k and subject s:
  - Draw distribution over subject-specific predicates $\xi_{k,s} \sim \text{Dir}(\gamma_s)$
- For each user u:
  - Draw user type index $z \sim \text{Cat}(\theta)$
  - For each proposition $\langle sbj, pred\rangle$ made by u:
    - Draw subject $sbj \sim \text{Cat}(\phi_z)$
    - Draw predicate $pred \sim \text{Cat}(\xi_{z,sbj})$

We set $K = 2$ in an attempt to recover a distinction between liberal and conservative users. For the experiments reported below, we run inference using collapsed Gibbs sampling (Griffiths and Steyvers, 2004) for 100 iterations, performing hyperparameter optimization on α, γ and γ_s (all asymmetric) every 10 using the fixed-point method of Minka (2003).

In order to compare the subject-specific predicate distributions across categories, we first calculate the posterior predictive distribution by taking a single sample of all latent variables z to estimate

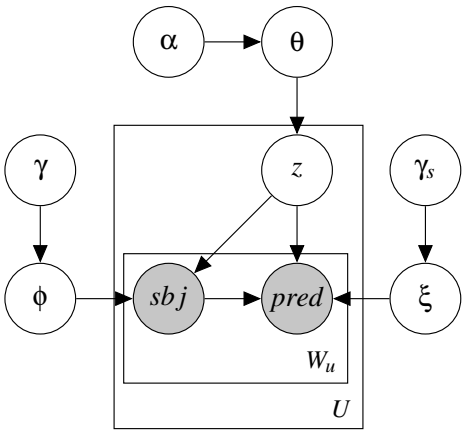

Figure 2: Single membership model with decoupled subjects and predicates. z is the latent category identity of a user (e.g., liberal or conservative); ϕ is a distribution over subjects for each category; and ξ is a distribution of predicates given subject s.

the following (Asuncion et al., 2009):

$$\hat{\zeta}_{z,v} = \frac{\mathbf{c}(z, v) + \gamma_v}{\sum_{v'} \mathbf{c}(z, v') + \gamma_{v'}} \tag{3}$$

Where $\hat{\zeta}_{z,v}$ is the vth element of the zth multinomial being estimated, $\mathbf{c}(z, v)$ is the count of element v associated with category z and γ_v is the associated Dirichlet hyperparameter for that element. Given this smoothed distribution, for each proposition we assign it a real valued score, the log-likelihood ratio between its value in these two distributions. In the evaluation that follows, let the PREDICATE SCORE for a given $\langle$subject, *predicate*$\rangle$ under this model be:

$$\log \left(\frac{\hat{\xi}_{0,\text{subject},\textit{predicate}}}{\hat{\xi}_{1,\text{subject},\textit{predicate}}} \right) \tag{4}$$

Let the PROPOSITION SCORE be:

$$\log \left(\frac{\hat{\phi}_{0,\text{subject}} \times \hat{\xi}_{0,\text{subject},\textit{predicate}}}{\hat{\phi}_{1,\text{subject}} \times \hat{\xi}_{1,\text{subject},\textit{predicate}}} \right) \tag{5}$$

4 Comparison

The two models described in §3 are unsupervised methods for estimating the latent political positions of users along with propositional assertions. We compare with three other models, a mixture of unsupervised, supervised, and semi-supervised methods. Unlike our models, these were not designed for the task described in §2.

4.1 Principal Component Analysis

To compare against another purely unsupervised model, we evaluate against principal component analysis (PCA), a latent linear model that minimizes the average reconstruction error between an original data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and a low-dimensional approximation $\mathbf{Z}\mathbf{W}^\top$, where $\mathbf{Z} \in \mathbb{R}^{n \times K}$ can be thought of as a K-dimensional latent representation of the input and $\mathbf{W} \in \mathbb{R}^{p \times K}$ contains the eigenvectors of the K largest eigenvalues of the covariance matrix $\mathbf{X}\mathbf{X}^\top$, providing a K-dimensional representation for each feature. We perform PCA with $K = 1$ on two representations of our data: a.) **counts**, where the input data matrix contains the counts for each proposition for each user, and b.) **frequencies**, where we normalize those counts for each user to unit length. While the input data is sparse, we must center each column to have a 0 mean (resulting in a dense matrix) and perform PCA through a singular value decomposition of that column-centered data using the method of Halko (2011); in using SVD for PCA, the right singular vectors correspond to the principal directions; from these we directly read off a $K = 1$ dimensional score for each proposition in our data.

4.2 ℓ_2-Regularized Logistic Regression

While unsupervised methods potentially allow us to learn interesting structure in data, they are often eclipsed in prediction tasks by the addition of any form of supervision. While purely supervised models give more control over the exact decision boundary being learned, they can suffer by learning from a much smaller training set than unsupervised methods have access to. To evaluate this tradeoff, we compare against a supervised model trained using naturally occurring data – users who self-declare themselves in their profiles to be *liberal*, *conservative*, *democrat*, or *republican*. We randomly sampled 150 users who self-identify as liberals and 150 who identify as conservatives. We do not expect these users to be a truly random sample of the population — those who self-declare their political affiliation are more likely to engage with political content differently from those who do not (Sandvig, 2015; Hargittai, 2015) — but is a method that has been used for political prediction tasks in the past (Cohen and Ruths, 2013).

We build a predictive model using two classes of features: a.) binary indicators of the most frequent 25,000 unigrams and multiword expressions[11] in the corpus overall; and b.) features derived from user posting activity to the seven blogs shown in table 1 (binary indicators of the blogs posted to, and the identity of the most frequent blog). In a tenfold cross-validation (using ℓ_2-regularized logistic regression), this classifier attains an accuracy rate of 76.7% (with a standard error of ± 1.7 across the ten folds).

In order to establish real-valued scores for propositions, we follow the same method as for the single membership model described above, using the log likelihood ratio of the probability of the proposition under each condition, where that probability is given as the count of the proposition among users classified as (e.g.) liberals (plus some small smoothing factor) divided by the total number of propositions used by them overall.

$$score(prop) = \log \frac{P(prop \mid z = conservative)}{P(prop \mid z = liberal)} \tag{6}$$

4.3 Co-Training

Since the features we use for the supervised model provide two roughly independent views of the data, we also evaluate against the semi-supervised method of co-training (Blum and Mitchell, 1998). Here, we train two different logistic regression classifiers, each with access to only the unigrams and multiword expressions employed by the user (h_{words}) or to binary indicators of the blogs posted to and the identity of the most frequent blog (h_{blogs}). For ten iterations, we pick a random sample U' of 1,000 data points from the full dataset U and classify each using the two classifiers; each classifier then adds up to 100 of the highest-confidence predictions to the training set, retaining the class distribution balance of the initial training set; after training, the final predictive probability for an item is the product of the two trained classifiers. In a tenfold cross-validation, co-training yielded a slightly higher (but not statistically significant) accuracy over pure supervision (77.0% ± 1.8). We calculate scores for propositions in the same way as for the fully supervised case above.

5 Evaluation

For the experiments that follow, we limit the input data available to all models to only those propo-

[11] Multiword expressions were found using the method of Justeson and Katz (1995).

sitions whose subject falls within the evaluation benchmark; and include only propositions used by at least five different users, and only users who make at least five different assertions, yielding a total dataset of 40,803 users and 1.9 million propositions (81,728 unique), containing the union of all three kinds of extracted propositions from §2.2.

Each of the automatic methods that we discuss above assigns a real-valued score to propositions like OBAMA IS A SOCIALIST. Our goal in evaluation is to judge how well those model scores recover those assigned by humans in our benchmark. Since each method may make different assumptions about the distribution of scores (and normalizing them may be sensitive to outliers), we do not attempt to model them directly, but rather use two nonparametric tests: Spearman's rank correlation coefficient and cluster purity.

Spearman's rank correlation coefficient. The set of scores in the human benchmark and as output by a model each defines a ranked list of propositions; Spearman's rank correlation coefficient (ρ) is a nonparametric test of the Pearson correlation coefficient measured over the ranks of items in two lists (rather than their values). We use the absolute value of ρ to compare the degree to which the ranked propositions of two lists are linearly correlated; a perfect correlation would have $\rho = 1.0$; no correlation would have $\rho = 0.0$.

Purity. While Spearman's rank correlation coefficient gives us a nonparametric estimate of the degree to which the exact order of two sequences are the same, we can also soften the exact ordering assumption and evaluate the degree to which a ranked proposition falls on the correct side of the political continuum (i.e., not considering whether OBAMA IS A SOCIALIST is more or less conservative than OBAMA IS A DICTATOR but rather that it is more conservative than liberal). For each ranked list, we form two clusters of propositions, split at the midpoint: all scores below the midpoint define one cluster, and all scores above or equal define a second. For $N = 766$ propositions, given gold clusters $\mathcal{G} = \{g_1, g_2\}$ and model clusters $\mathcal{C}_n = \{c_1, c_2\}$ (each containing 383 propositions), we calculate purity as the average overlap for the best alignment between the two gold clusters and their model counterparts.[12]

$$\text{Purity} = \frac{1}{N}\left(\max_j |g_1 \cap c_j| + \max_j |g_2 \cap c_j|\right) \tag{7}$$

A perfect purity score (in which all items from each cluster in $\mathcal{C}$ are matched to the same cluster in $\mathcal{G}$) is 1.0; given that all clusters are identically sized (being defined as the set falling on each half of a midpoint), a random assignment would yield a score of 0.50 in expectation.

Model	Purity	Spearman's ρ
Additive (PROP.)	0.757 $\pm$0.020	0.648 $\pm$0.017
Single mem. (PROP.)	0.754 $\pm$0.019	0.628 $\pm$0.017
Single mem. (PRED.)	0.702 $\pm$0.018	0.555 $\pm$0.015
Additive (PRED.)	0.705 $\pm$0.018	0.490 $\pm$0.013
Co-training	0.695 $\pm$0.018	0.450 $\pm$0.013
LR	0.619 $\pm$0.016	0.278 $\pm$0.010
PCA (frequency)	0.518 $\pm$0.014	0.098 $\pm$0.009
PCA (counts)	0.514 $\pm$0.014	0.066 $\pm$0.008

Table 3: Evaluation. Higher is better.

Table 3 presents the results of this evaluation. For both of the models described in §3, we present results for scoring a proposition like OBAMA IS A SOCIALIST based only on the conditional predicate score (PRED.) and on a score that includes variation in the subject as well (PROP.). Since both models are fit using approximate inference with a non-convex objective function, we run five models with different random initializations and present the average across all five.

We estimate confidence intervals using the block jackknife (Quenouille, 1956; Efron and Stein, 1981), calculating purity and Spearman's ρ over 76 resampled subsets of the full 766 elements, each leaving out 10.[13] For both metrics, the two best performing models show statistically significant improvement over all other models, but are not significantly different from each other.

We draw two messages from these results:

For heavily partisan data, unsupervised methods are sufficient. In drawing on comments on politically partisan blogs, we are able to match human judgments of the political import of propositions quite well (both of the unsupervised models

[12] In this case, with two clusters on each side, the best alignment in maximal in that $g_{n,i} \rightarrow c_{n,j} \Rightarrow g_{n,\neg i} \rightarrow c_{n,\neg j}$.

[13] As a clustering metric, purity has no closed-form expression for confidence sets, and since its evaluation requires its elements to be unique (in order to be matched across clusters), we cannot use common resampling-with-replacement techniques such as the bootstrap (Efron, 1979).

described in §3 outperform their supervised and semi-supervised counterparts by a large margin), which suggests that the easiest structure to find in this particular data is the affiliation of users with their political ideologies. Both unsupervised models are able to exploit the natural structure without being constrained by a small amount of training data that may be more biased (e.g., in its class balance) than helpful. The two generative models also widely outperform PCA, which may reflect a mismatch between its underlying assumptions and the textual data we observe; PCA treats data sparsity as structural zeros (not simply missing data) and so must model not only the variation that exists between users, but also the variation that exists in their frequency of use; other latent component models may be a better fit for this kind of data.

Joint information is important. For both models, including information about the full joint probability of a subject and predicate together yields substantial improvements for both purity and the Spearman correlation coefficient compared to scores calculated from variation in the conditional predicate alone. While we might have considered variation in the predicate to be sufficient in distinguishing between political parties, we see that this is simply not the case; variation in the subject may help anchor propositions in the spectrum relative to each other.

6 Convergent Validity

The primary quantity of interest that we are trying to estimate in the models described above is the political position of an *assertion*; a user's latent political affiliation is only a helpful auxiliary variable in reaching this goal. We can, however, also measure the correlation of those variables themselves with other variables of interest, such as users' self-declarations of political affiliation and audience participation on the different blogs. Both provide measures of convergent validity that confirm the distinction being made in our models is indeed one of political ideology.

6.1 Correlation with Self-declarations

One form of data not exploited by the unsupervised models described above are users' self-declarations; we omit these above in order to make the models as general as possible (requiring only text and not metadata), but they can provide an independent measure of the distinctions our unsupervised models are learning. (The supervised baselines in contrast are able to draw on this profile information for training data.)

Approximately 12% of the users in the data input to our models (4,718 of 40,804) have affiliated self-declared profile information; the most frequent of these include *retired*, *businessman*, *student*, and *patriot*. For all of these users, we regress binary indicators of the top 25,000 unigrams in their profiles against the MAP estimate of their political affiliation in the single-membership model. Across all 5 folds, the features with the highest predictive weights for one class were *patriot*, *conservative*, *obama*, and *god* while the highest predictive weights for the other are *progressive*, *voter*, *liberal*, and *science*.

6.2 Estimating Media Audience

We can also use users' latent political ideologies to estimate the overall ideological makeup of a blog's active audience. If we assign each post to our estimate of the political ideology of its author, we find that Mother Jones has the highest fraction of comments by estimated liberals at 80.4%, while Breitbart has the highest percentage of comments by conservatives (79.5%).

Blog	% Liberal by post
Mother Jones	80.4%
NPR	67.4%
Politico	51.6%
Young Conservatives	38.0%
Daily Caller	28.4%
IJ Review	28.0%
Breitbart	20.5%

Table 4: Media audience.

This broadly accords with Mitchell et al. (2014), which finds that among the blogs in our dataset, consistently liberal respondents trust NPR and Mother Jones most, while consistent conservatives trust Breitbart most and NPR and Mother Jones the least.

7 Conclusion

We introduce the task of estimating the political import of propositions such as OBAMA IS A SOCIALIST; while much work in open information extraction has focused on learning facts such as OBAMA IS PRESIDENT from text, we are able to exploit structure in the users and communities who make such assertions in order to align them all

within the same political space. Given sufficiently partisan data (here, comments on political blogs), we find that the unsupervised generative models presented here are able to outperform other models, including those given access to supervision.

One natural downstream application of this work is fine-grained opinion polling; while existing work has leveraged social media data on Twitter for uncovering correlations with consumer confidence, political polls (O'Connor et al., 2010), and flu trends (Paul and Dredze, 2011), our work points the way toward identifying fine-grained, interpretable propositions in public discourse and estimating latent aspects (such as political affiliation) of the communities who assert them. Data and code to support this work can be found at `http://people.ischool.berkeley.edu/~dbamman/emnlp2015/`.

8 Acknowledgments

We thank Jacob Eisenstein and our anonymous reviewers for their helpful comments. The research reported in this article was largely performed while both authors were at Carnegie Mellon University, and was supported by NSF grant IIS-1211277. This work was made possible through the use of computing resources made available by the Open Science Data Cloud (OSDC), an Open Cloud Consortium (OCC)-sponsored project.

References

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proc. of ICWSM*.

Arthur U. Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proc. of UAI*.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proc. of COLT*.

Robert Bond and Solomon Messing. 2015. Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review*, 109(01):62–78.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy! In *Proc. of ICWSM*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Bradley Efron and Charles Stein. 1981. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596.

Bradley Efron. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proc. of ICML*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proc. of EMNLP*.

Sean Gerrish and David M. Blei. 2012. How they vote: Issue-adjusted models of legislative behavior. In *NIPS*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5228–5235.

Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. 2011. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific Computing*, 33(5):2580–2594.

Eszter Hargittai. 2015. Why doesn't Science publish important methods info prominently? `http://goo.gl/wXUtys`, May.

Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson. 1993. Fastus: A system for extracting information from text. In *Proc. of HLT*.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proc. of ACL*.

Maojin Jiang and Shlomo Argamon. 2008. Exploiting subjectivity analysis in blogs to improve political leaning categorization. In *Proc. of SIGIR*.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(1):9–27.

Leslie Kaufman. 2014. Independent Journal Review website becomes a draw for conservatives. *New York Times*, Nov. 2, 2014.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of ACL*.

Andrew D. Martin and Kevin M. Quinn. 2002. Dynamic ideal point estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis*, 10(2):134–153.

Thomas P. Minka. 2003. Estimating a Dirichlet distribution. `http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/`.

Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. 2014. Political polarization and media habits: From Fox News to Facebook, how liberals and conservatives keep up with politics. Technical report, Pew Research Center.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proc. of AAAI*.

Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-aware truth assessment of fact candidates. In *ACL*.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. Tea party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th Congress. In *Proc. of ACL*.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. of ICWSM*.

Souneil Park, Minsam Ko, Jungwoo Kim, Ying Liu, and Junehwa Song. 2011. The politics of comments: Predicting political orientation of news stories with commenters' sentiment patterns. In *Proc. of CSCW*.

Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing twitter for public health. In *Proc. of ICWSM*.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, Republicans and Starbucks afficionados: User classification in Twitter. In *Proc. of KDD*.

Keith T. Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384.

Maurice H. Quenouille. 1956. Notes on bias in estimation. *Biometrika*, 43(3/4):353–360.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proc. of SMUC*.

Christian Sandvig. 2015. The Facebook "it's not our fault" study. `http://blogs.law.harvard.edu/niftyc/archives/1062`, May.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proc. of EMNLP*.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proc. of ACL*.

Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane J. Litman, David R. Pierce, Ellen Riloff, and Theresa Wilson. 2003. Recognizing and organizing opinions expressed in the world press. In *Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Theresa Ann Wilson. 2008. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states*. Ph.D. thesis, University of Pittsburgh.

Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. 2011. Classifying the political leaning of news articles and users from user votes. In *Proc. of ICWSM*.