An Analysis of Domestic Abuse Discourse on Reddit

Nicolas Schrading[1] **Cecilia O. Alm**[2] **Ray Ptucha**[1] **Christopher M. Homan**[3]
[1] Kate Gleason College of Engineering, Rochester Institute of Technology
[2] College of Liberal Arts, Rochester Institute of Technology
[3] Golisano College of Computing and Information Sciences, Rochester Institute of Technology
$\{jxs8172^{\S}|coagla^{\S}|rwpeec^{\S}|cmh^{\dagger}\}@\{^{\S}rit.edu|^{\dagger}cs.rit.edu\}$

Abstract

Domestic abuse affects people of every race, class, age, and nation. There is significant research on the prevalence and effects of domestic abuse; however, such research typically involves population-based surveys that have high financial costs. This work provides a qualitative analysis of domestic abuse using data collected from the social and news-aggregation website reddit.com. We develop classifiers to detect submissions discussing domestic abuse, achieving accuracies of up to 92%, a substantial error reduction over its baseline. Analysis of the top features used in detecting abuse discourse provides insight into the dynamics of abusive relationships.

1 Introduction

Globally, 30% of women fifteen and older have experienced physical and/or sexual intimate partner violence at some point in their life (Devries et al., 2013). While domestic abuse tends to have greater prevalence in low-income and non-western countries, it is still endemic in regions like North America and Western Europe. In the United States, by an intimate partner, 9.4% of women have been raped, 16.9% of women and 8% of men have experienced sexual violence other than rape, and 24.3% of women and 13.8% of men have experienced severe physical violence (Black et al., 2011). This translates to an estimated economic cost of $5.8 billion for direct medical and mental health care services, along with lost productivity and reduced lifetime earnings (Craft, 2003). Economic costs are calculable and provide concrete metrics for policy makers, but the physical and psychological effects felt by victims of domestic abuse are the true costs. Domestic abuse is the 12^{th} leading cause of years of life lost (Murray et al., 2013), and it contributes to health issues including frequent headaches, chronic pain, difficulty sleeping, anxiety, and depression (Black et al., 2011).

The data used to calculate such statistics are often derived from costly and time-consuming population-based surveys that primarily seek to obtain insight into the prevalence and consequences of domestic abuse. Due to safety concerns for victims and researchers, these surveys follow strict guidelines set by the World Health Organization (Garcia-Moreno et al., 2001). Great care must be taken by the researchers to ensure the safety of the participants, and therefore the number of participants is often quite small (Burge et al., 2014). One way to avoid the cost of wide scale surveys while still maintaining appropriate research conditions is to leverage the abundance of data publicly available on the web. Of particular interest are moderated forums that allow discourse between users.

Reddit[1] is one such website, and the chosen source of data for this paper. This site has a wide range of forums dedicated to various topics, called *subreddits*, each of which are moderated by community volunteers. For subreddits dedicated to sensitive topics such as depression, domestic abuse, and suicide, the moderators tend to ensure that the anonymous submitter has access to local help hotlines if a life-threatening situation is described. They also enforce respectful behavior and ensure that the submissions are on topic by deleting disrespectful or off-topic posts. Finally, they ensure that site rules are followed, including the strict disallowal of *doxing*, the practice of using submission details to reveal user identities.

Reddit allows lengthy submissions, unlike Twitter, and therefore the use of standard English is more common. This allows natural language processing tools like semantic role labelers trained on standard English to function better. Finally, Red-

[1] See `www.reddit.com`.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2577–2583,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

dit allows users to comment on submissions, providing them with the ability to ask questions, give advice, and provide support. This makes its data ideal for sensitive subjects not typically discussed in social media.

This work makes two contributions: classifiers for identifying texts discussing domestic abuse and an analysis of discussions of domestic abuse in several subreddits.

2 Related Work

Social media sites are an emerging source of data for public health studies, such as mental health, bullying, and disease tracking. These sites provide less intimidating and more accessible channels for reporting, collectively processing, and making sense of traumatic and stigmatizing experiences (Homan et al., 2014; Walther, 1996). Many researchers have focused on Twitter data, due to its prominent presence, accessibility, and the characteristics of tweets. For instance, De Choudhury et al. (2013) predicted the onset of depression from user tweets, while other studies have modeled distress (Homan et al., 2014; Lehrman et al., 2012). Most relevantly, Schrading et al. (2015) used the #WhyIStayed trend to predict whether a tweet was about staying in an abusive relationship or leaving, analyzing the lexical structures victims of abuse give for staying or leaving.

Reddit has been studied less in this area, with work mainly focusing on mental health. In Pavalanathan and De Choudhury (2015), a large number of subreddits on the topic of mental health were identified and used to determine the differences in discourse between throwaway[2] and regular accounts. They observed almost 6 times more throwaway submissions in mental health subreddits over control subreddits, and found that throwaway accounts exhibit considerable disinhibition in discussing sensitive aspects of the self. This motivates our work in analyzing Reddit submissions on domestic abuse, which can be assumed to have similar levels of throwaway accounts and discussion. Additionally, Balani and De Choudhury (2015) used standard ngram features, along with submission and author attributes to classify a submission as high or low self-disclosure.

3 Dataset[3] and Data Analysis

Following the procedure in Balani and De Choudhury (2015) for subreddit discovery, we identified several subreddits that focus on domestic abuse. Additionally, we determined subreddits unrelated to domestic abuse, to be used as a control set. Table 1 shows the subreddits, the total number of unique posts (called *submissions*[4]), and total number of replies to those submissions (called *comments*) collected.

Domestic Abuse	**# Submissions**	**# Comments**
abuseinterrupted	1653	1069
domesticviolence	749	2145
survivorsofabuse	512	2172
Control	**# Submissions**	**# Comments**
casualconversation	7286	285575
advice	5913	31323
anxiety	4183	23300
anger	837	3693

Table 1: The domestic abuse subreddits and control subreddits with the total number of submissions and comments collected.

The *anger* and *anxiety* subreddits were chosen as control subreddits in order to help the classifier discriminate between the dynamics of abusive relationships and the potential effects of abuse on victims. For example, anxiety and anger may be affect caused by domestic abuse, but they are also caused by a wide variety of other factors. By including these subreddits in the control set, a classifier should utilize the situations, causes, and stakeholders in abusive relationships as features, not the affect particularly associated with abusive relationships. Similarly, the *advice* subreddit was chosen as a way to help the classifier understand that advice-seeking behavior is not indicative of abuse. The *casualconversation* subreddit allows discussion of anything, providing an excellent sample of general written discourse. The domestic abuse subreddits have far fewer active users, submissions, and comments in total.

3.1 Preprocessing

All experiments used the same preprocessing steps. From the collected subreddits, only submissions with at least one comment were chosen to be included for study. We then ran the submission text through the Illinois Curator (Clarke et

[2] Anonymous, one-time accounts to submit a single (often personal or sensitive) submission or comment.

[3] Released on nicschrading.com/data/.

[4] Submission text is its title and selftext (an optional text body) concatenated together.

al., 2012) to provide semantic role labeling (SRL) (Punyakanok et al., 2008). A total of 552 domestic abuse submissions were parsed, and we randomly chose an even distribution of the control subreddits (138 each), yielding a total sample size of 1104. All submissions were normalized by lowercasing, lemmatizing, and stoplisting. External links and URLs were replaced with *url* and references to subreddits, e.g. */r/domesticviolence*, were replaced with *subreddit_link*.

3.2 Descriptive Statistics

We present basic descriptive statistics on the set of 552 abuse submissions and 552 non-abuse submissions in Table 2.

	Abuse	**Non-Abuse**
Avg comments/post	5.4 ± 6.1	13.2 ± 25.3
Avg score/post	6.1 ± 5.1	7.5 ± 16.4
Avg tokens/post	278 ± 170	208 ± 164
# unique submitters	482	535
Avg comment depth	0.96 ± 1.5	1.5 ± 1.9
Avg comment score	2.2 ± 2.7	2.1 ± 2.9
Avg tokens/comment	107 ± 128	53.4 ± 79.9
# comments	2989	6964
# unique commenters	1022	2519

Table 2: Basic descriptive statistics. The score is provided by users voting on submissions/comments they feel are informative. The depth of a comment indicates where in a reply chain it falls. A depth of 0 means it is in reply to the submission, a depth of 1 means it is in reply to a depth 0 comment, etc. The ± values are standard deviation metrics.

In general, the non-abuse subreddits have more discourse between commenters, as indicated by a larger comment depth, however, the abuse subreddits tend to have longer submissions and replies. The abuse subreddits perhaps also have a smaller more tight-knit, community as indicated by fewer numbers of unique submitters and commenters.

3.3 Ngram Attributes

To get a sense of the language used between the two sets of subreddits, the most frequent 1-, 2-, and 3-grams were examined. While there are many common and overlapping ngrams in the two sets, each set does have distinct ngrams. In the abuse set, distinct ngrams include the obvious *abuse* (1595 occurences), *domestic violence* (202), and *abusive relationship* (166). Additionally, unique 3-grams related to the agents and situations in abusive relationships like *local dv agency* (12) and *make feel bad* (11) appear. Also included are unique empathetic and helping discourse from comments, including *let know* (121), and *feel free pm*[5] (27). This indicates that comment data could improve classification results, as support and empathy may be more prevalent in the *abuse* set than in the control set.

3.4 Semantic Role Attributes

From the SRL tool, our dataset was tagged with various arguments of predicates. This data is particularly useful in our study, as we are interested in examining the semantic actions and stakeholders within an abusive relationship. By performing a lookup in Proposition Bank (Martha et al., 2005) with a given argument number, predicate, and sense, we retrieved unique role labels for each argument.

We determined the top 100 most frequent roles and predicates in the two sets, and took only the unique roles and predicates within each set to see what frequently occurring but unique roles and predicates exist within the abuse and control group.

Role Label	**Predicate**
caller, 175	abuse, 433
thing hit, 174	share, 167
agent, hitter - animate only!, 164	believe, 164
abuser, agent, 162	call, 151
entity abused, 139	remember, 149
utterance, 115	cry, 147
patient, entity experiencing hurt/damage, 113	!tell, 142
utterance, sound, 104	send, 127
belief, 104	thank, 127
benefactive, 103	realize, 124

Table 3: Top 10 unique roles and predicates with their frequency for the abuse data. An exclamation point on a predicate indicates negation.

Table 3 contains roles and predicates that are powerful indicators of an abusive relationship, including a *caller*, *hitter*, *thing hit*, *abuser*, and *entity experiencing hurt/damage*. Importantly, several predicates that appear in this data also appear in a study on discussions of domestic abuse in Twitter data, including *believe* and *realize*, which indicate cognitive manipulation in the victims of domestic abuse (Schrading et al., 2015).

[5] The initials *pm* stand for private message.

Classifer	N	P	R	N+P	N+R	P+R	N+P+R
Linear SVM	**90 ± 3**	72 ± 5	73 ± 4	88 ± 3	88 ± 3	73 ± 4	87 ± 3

Table 4: Classification accuracies of Linear SVM. N=Ngrams, P=Predicates, R=Roles.

4 Classification Experiments

In order to discover the semantic and lexical features salient to abusive relationships, we designed several classifiers. The subreddit category to which a submission was posted was used as the gold standard label of *abuse* or *non-abuse*. The labels were validated by examining the top ngrams, roles, and predicates in Section 3, and taking into account that these subreddits are moderated for on-topic content. We ran several experiments to study classifiers, the impact of features, and the effect of comments on prediction performance.

4.1 Combinations of Features

We used the 1-, 2-, and 3-grams in the submission text, the predicates, and the semantic role labels as features, using TF-IDF vectorization[6]. Perceptron, naïve Bayes, logistic regression, random forest, radial basis function SVM, and linear SVM classifiers were parameter optimized using 10-fold cross validation. Table 4 contains the results for the best classifier. The best features are the ngrams, achieving the highest accuracies alone. Predicates and semantic roles perform admirably, but bring the classifier accuracies down slightly when added to ngrams. To determine the top features for prediction, we examined the weights of the top performing classifier, Scikit-learn's (Pedregosa et al., 2011) Linear SVM with C=0.1, as in Guyon et al. (2002). These, along with their weights, are shown in Table 5.

Abuse	Non Abuse
abusive, 1.3	anxiety, 1.1
child, 0.93	anger, 1.1
abuser, 0.86	job, 0.52
relationship, 0.84	school, 0.46
therapy, 0.83	hour, 0.45
survivor, 0.83	week, 0.45
domestic, 0.73	fuck, 0.44
happen, 0.72	class, 0.42
violence, 0.68	college, 0.41
father, 0.67	fun, 0.40

Table 5: Top 10 features based on Linear SVM weights using only ngrams from submissions. The classifier may be relying heavily on the anxiety and anger subreddits to discriminate between abuse and non-abuse, as indicated by the sharp drop in SVM weight from *anger* to *job*. Abuse word weights are more evenly distributed.

4.2 Comment Data Only

We experimented with only comment data to predict if they were posted in an *abuse* or *non-abuse* subreddit. Because ngram features performed best in the previous experiment, a larger set of submissions (1336 per class) was used. A final held out testset was created from 10% of these submissions, giving 1202 submissions per class for the devset and 134 per class for the testset. Taking the comments from these submissions yielded 4712 abuse and 19349 non-abuse comments for the devset and 642 abuse and 2264 non-abuse comments for the testset. 10-fold cross-validation was used on the devset to tune the classifier. Using a Linear SVM with C=1 achieved an F1 score of $0.70 \pm .02$ on the devset. On the held out testset, it achieved a precision of 0.68, recall of 0.62, and F1 score of 0.65. Examining its weights gives features similar to those in Table 5, with additional empathetic discourse like *thank*, *hug*, and *safe* in the abuse class.

4.3 Comment and Submission Text Combined

Concatenating the comments to their respective submissions may improve results, but because comments can be completely off-topic or in reply to other comments, we experimented with only the top-scoring comments and those most similar to the submission text. To compute similarity we used a sum of the word vector representations from Levy and Goldberg (2014) as included in spaCy (Honnibal, 2015) and used cosine similarity. Taking only the top 90^{th} percentile in user voting score and text similarity, we had 2688 abuse comments and 7852 non-abuse comments concatenated to the 1336 submissions of their class. This method achieves extremely high accuracy of $94\% \pm 2\%$ on the devset and 92% on the testset using a Linear SVM with C=1. A classifier trained only on the submission text data from the same devset/testset split obtains the lower accuracies of

[6] Binary features and only unigrams were tried but these did not improve results.

90%±2% on the devset and 86% on the testset, indicating that comments can add predictive power. The top features are similar to those in Table 5.

4.4 Uneven Set of Submissions

Using the method in Section 4.3 to train the classifier, a larger, uneven set of data was examined (still using only ngrams). This set contained 1336 *abuse* and 17020 *non-abuse* instances. From this set, 15% were held out for final examination as a testset and the rest was used as a devset with 5-fold cross-validation. On the devset, an F1 score of 0.81 ± 0.01 was achieved while on the testset it had a precision of 0.84, recall of 0.74, and F1 score of 0.79. The best classifier was a Linear SVM with C=100. The confusion matrix of the testset is in Table 4.4.

		Predicted Class	
		Abuse	**Non-Abuse**
Actual Class	**Abuse**	152	53
	Non-Abuse	29	2520

Table 6: Confusion matrix on the testset of the Abuse/Non-Abuse classifier.

This classifier has good precision for the *abuse* class, and decent recall, meaning that there can be confidence that submissions flagged as *abuse* are indeed about *abuse*. By applying this classifier to a large held out set of data, these results suggest that many potentially relevant submissions would be flagged for examination, and they would mostly be about *abuse*.

4.5 Testing on Completely Held Out Subreddits

To get a sense of efficacy in the wild in detecting submissions about abuse, the best classifier from Section 4.4 was taken (trained on the devset data) and run on a large set of submissions from the *relationships* and *relationship_advice* subreddits. These subreddits are general forums for discussion and advice on any relationship (not necessarily intimate). Their submissions tend to be long, descriptive, and extremely personal.

After running the abuse classifier on the submissions from these subreddits with at least 1 comment (13623 in total, with their 90^{th} percentile comments concatenated), 423 submissions were flagged as being about abuse. 101 of these 423 were annotated by 3 annotators (co-authors), using the labels *A* (the submission discusses an abusive relationship), *M* (off-hand mention of abuse), *N* (not about abuse), and *O* (off-topic submission or other).

From the three annotators' annotations, on average 59% are *A*, 16% are *M*, 23% are *N*, and 2% are *O*. The percentage of overall agreement was 72% and Randolph's free-marginal multirater kappa (Warrens, 2010) score was 0.63. Annotators occasionally had a hard time distinguishing between *A* and *M*, as context may have been missing. Combining the two by considering all *M* as *A*, the average percent of *A* increases to 75%, the percentage of overall agreement improves to 86% and the kappa score improves to 0.79. Taking the statistic that on average 75% of the flagged submissions in the annotated subset are about abuse or have a mention of abuse indicates that this classifier should hopefully have a precision of around 0.75 on unseen Reddit data at large. Understandably, the precision drops by about .1 compared to its use on the subreddits it was trained and tested on. A precision of 0.75 on this set of data would mean that any statistics from this set may include some noise, but overall, the trends should reveal important results about abuse.

5 Conclusion

This work provides an analysis of domestic abuse using the online social site Reddit. Language analysis reveals interesting patterns used in discussing abuse, as well as initial data about the semantic actions and stakeholders involved in abusive relationships. Multiple classifiers were implemented to determine the top semantic and linguistic features in detecting abusive relationships. Simpler features such as ngrams performed above the more complex predicate and role labels extracted from a semantic role labeler, though the more complex structures contribute to interesting insights in data analysis. Future work could use a larger training set from multiple online sites to analyze the patterns of online abuse discourse across varied forums.

Acknowledgement

This work was supported in part by a GCCIS Kodak Endowed Chair Fund Health Information Technology Strategic Initiative Grant and NSF Award #SES-1111016.

References

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual Association for Computing Machinery Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1373–1378, New York, NY, USA. ACM.

Michele C Black, Kathleen C Basile, Matthew J Breiding, Sharon G Smith, Mikel L Walters, Melissa T Merrick, and MR Stevens. 2011. National intimate partner and sexual violence survey. *Atlanta, GA: Centers for Disease Control and Prevention*, 75.

Sandra K. Burge, Johanna Becho, Robert L. Ferrer, Robert C. Wood, Melissa Talamantes, and David A. Katerndahl. 2014. Safely examining complex dynamics of intimate partner violence. *Families, Systems, & Health*, 32(3):259 – 270.

Munmun De Choudhury, Scott Counts, Eric Horvitz, and Michael Gamon. 2013. Predicting depression via social media. In *Proceedings of the Seventh Annual International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, pages 128–137. AAAI, July.

James Clarke, Vivek Srikumar, Mark Sammons, and Dan Roth. 2012. An NLP curator (or: How I learned to stop worrying and love NLP pipelines). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3276–3283, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Carole Craft. 2003. Costs of intimate partner violence against women in the United States. Technical report, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta, GA.

K.M. Devries, Joelle Y.T. Mak, C. García-Moreno, M. Petzold, J.C. Child, G. Falder, S. Lim, L.J. Bacchus, R.E. Engell, L. Rosenfeld, C. Pallitto, T. Voss, and C.H. Watts. 2013. The global prevalence of intimate partner violence against women. *Science*, 340(6140):1527–1528.

Claudia Garcia-Moreno, C Watts, and L Heise. 2001. Putting women first: Ethical and safety recommendations for research on domestic violence against women. *Department of Gender and Womens Health, World Health Organization. Geneva, Switzerland.*

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, March.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Matthew Honnibal. 2015. SpaCy: Industrial strength NLP with Python and Cython. `https://github.com/honnibal/spaCy`.

Michael Thaul Lehrman, Cecilia Ovesdotter Alm, and Rubén A. Proaño. 2012. Detecting distressed and non-distressed affect states in short forum texts. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 9–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Palmer Martha, Gildea Dan, and Kingsbury Paul. 2005. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31:71–106.

Christopher JL Murray, Jerry Abraham, Mohammed K Ali, Miriam Alvarado, Charles Atkinson, Larry M Baddour, David H Bartels, Emelia J Benjamin, Kavi Bhalla, Gretchen Birbeck, et al. 2013. The state of US health, 1990-2010: Burden of diseases, injuries, and risk factors. *JAMA*, 310(6):591–606.

Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity management and mental health discourse in social media. In *Proceedings of WWW'15 Companion: 24th International World Wide Web Conference, Web Science Track*, Florence, Italy, May. WWW'15 Companion.

Fabian Pedregosa, Gaël Varoquaux., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287.

Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. 2015. #WhyIStayed, #WhyILeft: Microblogging to make sense of domestic abuse. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1281–1286, Denver, Colorado, May–June. Association for Computational Linguistics.

Joseph Walther. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1):3–43, Feb.

Matthijs J. Warrens. 2010. Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, 4(4):271–286.