#WhyIStayed, #WhyILeft: Microblogging to Make Sense of Domestic Abuse

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

In September 2014, Twitter users unequivocally reacted to the Ray Rice assault scandal by unleashing personal stories of domestic abuse via the hashtags `#WhyIStayed` or `#WhyILeft`. We explore at a macro-level firsthand accounts of domestic abuse from a substantial, balanced corpus of tweeted instances designated with these tags. To seek insights into the reasons victims give for staying in vs. leaving abusive relationships, we analyze the corpus using linguistically motivated methods. We also report on an annotation study for corpus assessment. We perform classification, contributing a classifier that discriminates between the two hashtags exceptionally well at 82% accuracy with a substantial error reduction over its baseline.

1 Introduction

Domestic abuse is a problem of pandemic proportions; nearly 25% of females and 7.6% of males have been raped or physically assaulted by an intimate partner (Tjaden and Thoennes, 2000). These numbers only include physical violence; psychological abuse and other forms of domestic abuse may be even more prevalent. There is thus an urgent need to better understand and characterize domestic abuse, in order to provide resources for victims and efficiently implement preventative measures.

Survey methods exploring domestic abuse involve considerable time and investment, and may suffer from under-reporting, due to the taboo and stressful nature of abuse. Additionally, many may not have the option of directly seeking clinical help. Social media may provide a less intimidating and more accessible channel for reporting, collectively processing, and making sense of traumatic and stigmatizing experiences (Homan et al., 2014; Walther, 1996). Such data has been used for analyzing and predicting distinct societal and health issues, aimed at improving the understanding of wide-reaching societal concerns. For instance, 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). Xu et al. (2013) used Twitter data to identify bullying language, then analyzed the characteristics of these tweets, and forecasted if a tweet would be deleted out of regret.

In September 2014, in the wake of the Ray Rice assault scandal[1] and the negative public reaction to the victim's decision to stay and support her abuser, Twitter users unequivocally reacted in a viral discussion of domestic abuse, defending the victim using the hashtag #WhyIStayed and contrasting those with #WhyILeft. Such narrative sharing may have a cathartic and therapeutic effect, extending the viral reach of the trend.

Analysis of the linguistic structures embedded in these tweet instances provides insight into the critical reasons that victims of domestic abuse report for choosing to stay or leave. Trained classifiers agree with these linguistic structures, adding evidence that these social media texts provide valuable insights into domestic abuse.

[1] `http://www.sbnation.com/nfl/2014/5/23/5744964/ray-rice-arrest-assault-statement-apology-ravens`.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1281–1286,
Denver, Colorado, May 31 – June 5, 2015.

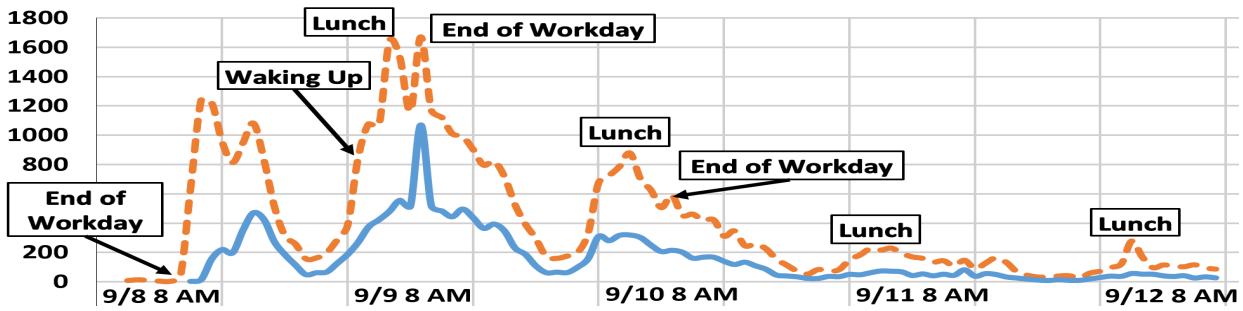

Figure 1: Tweet count per hour with #WhyIStayed (dotted) or #WhyILeft (solid) from 9/8 to 9/12. Times in EST, vertical lines mark 12 hour periods, with label corresponding to its left line. Spam removed, includes meta tweets.

2 Data

We collected a new corpus of tweets using the Twitter and Topsy[2] application programming interfaces. The corpus spans the beginning of September (the start of the trend) to the beginning of October, 2014. We fully rehydrated the tweets (to update the retweet count, etc.) at the end of the collection period. Figure 1 displays the behavior from the initial days of this trend. Due to its viral nature, the majority of tweets are from the first week of the trend's creation.

2.1 Preprocessing

We removed spam tweets based on the usernames of the most prevalent spammers, as well as key spam hashtags.[3] We also removed tweets related to a key controversy, in which the Twitter account for DiGiorno Pizza (ignorant of the trend's meaning) tweeted *#WhyIStayed You had pizza*.[4] This resulted in over 57,000 unique tweets in the corpus.

Many tweets in the dataset were reflections on the trend itself or contained messages of support to the users sharing their stories, for example, *Not usually a fan of hashtag trends, but #WhyIStayed is incredibly powerful. #NFL #RayRice*.[5] These tweets, here denoted *meta-tweets*, were often retweeted, but they rarely contained reasons for staying or leaving (our interest), so we filtered them out by keyword.[6] In section 2.3 we empirically explore the remaining instances.

[2]For outside Twitter's history, `http://topsy.com/`

[3]Such as *#MTVEMA, #AppleWatch, #CMWorld*.

[4]Removed by keywords *pizza, digiorno*.

[5]Illustrative tweet examples were anonymized and we purposefully attempted to minimize inclusion of sensitive content.

[6]Including *janay/ray rice, football, tweets, trend, video, etc.*

2.2 Extracting Gold Standard Labels

Typically, users provided reasons for staying and leaving, with the reasons prefixed by or appended with the hashtags #WhyIStayed or #WhyILeft as in this example: *#WhyIStayed because he told me no one else would love me. #WhyILeft because I gained the courage to love myself.* Regular expressions matched these structures and for tweets marked by both tags, split them into multiple instances, labeled with their respective tag. If the tweet contained only one of the target hashtags, the instance was labeled with that hashtag. If the tweet contained both hashtags but did not match with any of the regular expressions, it was excluded to ensure data quality.

The resulting corpus comprised 24,861 #WhyIStayed and 8,767 #WhyILeft labeled datapoints. The class imbalance may be a result of the origins of the trend rather than an indicator that more victims stay than leave. The tweet that started the trend contained only the hashtag #WhyIStayed, and media reporting on the trend tended to refer to it as the "#WhyIStayed phenomenon." As Figure 1 shows, the first #WhyILeft tweet occurred hours after the #WhyIStayed trend had taken off, and never gained as much use. By this reasoning, we concluded that an even set of data would be appropriate, and enable us to use the ratio metric in experiments discussed in this paper, as well as compare themes in the two sets. By random sampling of #WhyIStayed, a balanced set of 8,767 examples per class was obtained, resulting in a binary 50% baseline. From this set, 15% were held out as a final testset, to be considered after a tuning procedure with the remaining 85% devset.

2.3 Annotation Study

Four people (co-authors) annotated a random sample of 1000 instances from the devset, to further characterize the filtered corpus and to assess the automated extraction of gold standard labels. This random subset is composed of 47% #WhyIStayed and 53% #WhyILeft gold standard samples. Overall agreement overlap was 77% and Randolph's free-marginal multirater kappa (Warrens, 2010) score was 0.72. According to the annotations in this random sample, on average 36% of the instances are reasons for staying (S), 44% are reasons for leaving (L), 12% are meta comments (M), 2% are jokes (J), 2% are ads (A), and 4% do not match prior categories (O). Table 1 shows that most related directly to S or L, with annotators identifying more clearly L. Of interest are examples in which annotators did not agree, as these are indicative of problems in the data, and are samples that a classifier will likely label incorrectly. The tweet *because i was slowly dying anyway* was marked by two annotators as S and two annotators as L. Did the victim have no hope left and decide to stay? Or did the victim decide that since they were "slowly dying anyway" they could attempt to leave despite the possibility of potentially being killed in the attempt? The ground truth label is #WhyILeft. Another example with two annotators labeling as S and two as L is *two years of bliss, followed by uncertainty and fear*. This tweet's label is #WhyIStayed. The limited context from these samples makes it difficult to interpret fully, and causes human annotators to fail; however, most cases contain clear enough reasoning to interpret correctly.

		A	J	L	M	O	S
A1	#L	.01	.01	.78	.11	.03	.07
	#S	.01	.03	.10	.21	.02	.63
A2	#L	.02	.01	.72	.06	.09	.10
	#S	.03	.01	.07	.16	.10	.63
A3	#L	.00	.02	.77	.09	0	.11
	#S	.01	.04	.06	.21	0	.68
A4	#L	.02	.01	.75	.05	.04	.14
	#S	.03	.01	.16	.12	.05	.63

Table 1: Confusion matrices of all 4 annotators, compared to the gold standard. Annotators mostly identified reasons for staying or leaving, and only a small fraction were unrelated. #L=#WhyILeft, #S=#WhyIStayed.

3 Methods for Exploring Reasons

3.1 Cleaning and Classifier Tuning

All experiments used the same cleaned data: removing hashtags, replacing URLs with the token *url* and user mentions with *@mention*, and replacing common emoticons with a sentiment indicator: *emotsent*$\{p|n|neut\}$ for positive/negative/neutral. Informal register was expanded to standard English forms using a slang dictionary.[7] Classifier tuning involved 5-fold cross-validation and selecting the best parameters based on the mean accuracy. For held-out data testing the full devset was used for training.

3.2 Analysis of Vocabulary

We examined the vocabulary in use in the data of the two hashtag sets by creating a frequency distribution of all unigrams after stoplisting and lowercasing. The wordcloud unigrams in Figure 2 are weighted by their relative frequency. These wordclouds hint at the reasons; however, decontextualized unigrams lead to confusion. For example, why does *left* appear in both? Other experiments were done to provide context and expand analysis.

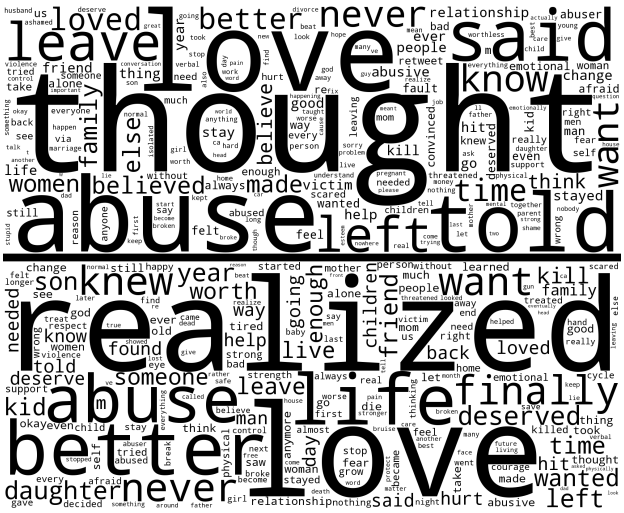

Figure 2: A wordcloud of unigrams, weighted by unigram frequencies, for (top) #WhyIStayed instances and (bottom) #WhyILeft instances.[8]

[7] http://www.noslang.com/

[8] Created using http://amueller.github.io/word_cloud/

Most discriminative *abuser onto victim* verbs									Legend
convince	find	isolate	kick	kill	love	manipulate	promise	want	#WhyIStayed
0.96	1	0.93	1	0.91	0.95	1	0.83	0.93	#WhyILeft
Most discriminative *victim as subject* verbs									
believe	choose	decide	felt	know	learn	realize	think	want	
0.81	1	1	0.79	0.82	1	0.99	0.93	0.83	

Table 2: Discriminative verbs for *abuser onto victim* and *victim as subject* structures.

3.3 Analysis of Subject-Verb-Object Structures

Data inspection suggested that many users explained their reasons using a Subject-Verb-Object (SVO) structure, in which the abuser is doing something to the victim, or the victim is explaining something about the abuser or oneself.[9] We used the open-source tools Tweeboparser (Kong et al., 2014) and TurboParser (Martins et al., 2013) to heuristically extract syntactic dependencies, constrained by pronomial usage. Both parsers performed similarly, most likely due to the well-formed English in the corpus. While tweets are known for non-standard forms, the seriousness of the discourse domain may have encouraged more standard writing conventions.

Using TurboParser, we conducted an analysis for both male and female genders acting as the abuser in the subject position. Starting at the lemmatized predicate verb in each dependency parse, if the predicate verb followed an abuser subject word[10] per the dependency links, and preceded a victim object word,[11] it was added to a conditional frequency distribution, with the two classes as conditions. These structures are here denoted *abuser onto victim*. We used similar methods to extract structures in which the victim is the subject. Instances with female abusers were rare, and statistical gender differences could not be pursued. Accordingly, both genders' frequency counts were combined. Discriminative predicates from these conditional frequency distributions were determined by equation (1). In Table 2 we report on those where the ratio is greater than 0.75 and the total count exceeds a threshold to avoid bias towards lower frequency verbs.

$$ratio = \frac{count_{largerOfCounts}}{count_{left} + count_{stayed}} \quad (1)$$

[9]Example: *He hurt my child* S: *He*, V: *hurt*, O: *my child*.

[10]Male abuser: *he, his, my bf*, etc. Female: *she, her*, etc.

[11]Male victim: *me, my, him*, etc. *Female: me, my, her*, etc.

3.4 Classification Experiments

We examined the usefulness of the SVO structures, using subsets of the devset and testset having SVO structures (10% of the instances in total). While 10% is not a large proportion overall, given the massive number of possible dependency structures, it is a pattern worth examining – not only for corpus analytics but also classification, particularly as these SVO structures provide insight into the abuser-victim relationship. A linear SVM using boolean SVO features performed best (C=1), obtaining $70\% \pm 2\%$ accuracy on the devset and 73% accuracy on the testset. The weights assigned to features by a Linear SVM are indicative of their importance (Guyon et al., 2002). Here, the top features presented as (S,V,O) for #WhyIStayed were: *(he, introduce, me)*, *(i, think, my)*, *(he, convince, me)*, *(i, believe, his)*, and *(he, beat, my)*. For #WhyILeft they were *(he, choke, me)*, *(i, beg, me)*, *(he, want, my)*, *(i, realize, my)*, and *(i, listen, my)*.

The SVO structures capture meaning related to staying and leaving, but are limited in their data coverage. Another experiment explored an extended feature set including uni-, bi-, and trigrams in sublinear $tf \times idf$ vectors, tweet instance character length, its retweet count, and SVO structures. We compared Naïve Bayes, Linear SVM, and RBF SVM classifiers from the Scikit-learn package (Pedregosa et al., 2011). The RBF SVM performed slightly better than the others, achieving a maximum accuracy of $81\% \pm .3\%$ on the devset and 82% on the testset.[12,13] Feature ablation, following the procedure in Fraser et al. (2014), was utilized to determine the most important features for the classifier, the results

[12]Tuned parameters: max df = 11%, C=10, gamma=1.

[13]Dimensionality reduction with Supervised Locality Preserving Projections (SLPP) (Ptucha and Savakis, 2014) was attempted, but this did not improve results.

of which can be seen in Table 3.

Removed	Remaining Features	% Acc
	NG+E+IR+TL+RT+SVO	81.90
SVO	NG+E+IR+TL+RT	82.09
TL	NG+E+IR+RT	82.21
E	NG+IR+RT	82.21
RT	NG+IR	82.13
IR	NG	81.48

Table 3: Feature ablation study with an RBF SVM and no dimensionality reduction. NG = ngrams, E = emoticon replacement, IR = informal register replacement, TL = tweet length, RT = retweet count, SVO = subject-verb-object structures. % Acc is accuracy on the testset.

Interestingly, the SVO features combined with ngrams worsened performance slightly, perhaps due to trigrams capturing the majority of SVO cases. The highest accuracy, 82.21% on the testset, could be achieved with a combination of ngrams, informal register replacement, and retweet count. However the vast majority of cases can be classified accurately with ngrams alone. Emoticons may not have contributed to performance since they were rare in the corpus. Standardizing non-standard forms presumably helped the SVM slightly by boosting the frequency counts of ngrams while removing non-standard ngrams. Tweet length reduced accuracy slightly, while the number of retweets helped.

4 Discussion

From the analyses of SVO structures, word-clouds, and Linear SVM weights, interesting micro-narratives of staying and leaving emerge. Victims report staying in abusive relationships due to cognitive manipulation, as indicated by a predominance of verbs including *manipulate, isolate, convince, think, believe, felt* while report leaving when experiencing or fearing physical violence, via predicates such as *kill* and *kick*. They also report staying when in dire financial straits (*money*), when attempting to keep the nuclear family united (*family, marriage*) or when experiencing shame about their situation (*ashamed, shame*). They report leaving when threats are made towards loved-ones (*son, daughter*), gain agency (*choose, decide*), realize their situation or self-worth (*realize, learn, worth, deserve, finally, better*), or gain support from friends or family (*courage, support, help*). Importantly, such reasons for staying are validated in the clinical literature (Buel, 1999).

5 Conclusion

We discuss and analyze a filtered, balanced corpus having the hashtags #WhyIStayed or #WhyILeft. Our analysis reveals micro-narratives in tweeted reasons for staying vs. leaving. Our findings are consistent across various methods, correspond to observations in the clinical literature, and affirm the relevance of NLP for exploring issues of social importance in social media. Future work will focus on improving SVO extraction, especially adding consideration for negations of predicate verbs. In addition we will analyse other hashtags in use in the trend and perform further analysis of the trend itself, implement advanced text normalization rather than relying on a dictionary, and determine the roles features from linked webpages and FrameNet or other semantic resources play in making sense of domestic abuse.

6 Acknowledgement

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

References

Sarah M. Buel. 1999. Fifty obstacles to leaving, a.k.a, why abuse victims stay. *The Colorado Lawyer*, 28(10):19–28, Oct.

Munmun De Choudhury, Scott Counts, Eric Horvitz, and Michael Gamon. 2013. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Cambridge, Massachusetts, July. Association for the Advancement of Artificial Intelligence.

Kathleen C. Fraser, Graeme Hirst, Naida L. Graham, Jed A. Meltzer, Sandra E. Black, and Elizabeth Rochon. 2014. Comparison of different feature sets for identification of variants in progressive aphasia. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 17–26, Baltimore, Mary-

land, USA, June. Association for Computational Linguistics.

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.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October. Association for Computational Linguistics.

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.

Andre Martins, Miguel Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria, August. Association for Computational Linguistics.

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.

Raymond Ptucha and Andreas Savakis. 2014. LGE-KSVD: Robust sparse representation classification. *IEEE Transactions on Image Processing*, 23(4):1737–1750, April.

Patricia Tjaden and Nancy Thoennes. 2000. Extent, nature, and consequences of intimate partner violence: Findings from the national violence against women survey. Technical Report NCJ 181867, National Institute of Justice, Centers for Disease Control and Prevention, Washington, DC.

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.

Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. 2013. An examination of regret in bullying tweets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 697–702, Atlanta, Georgia, June. Association for Computational Linguistics.