The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Extract Semantic Pattern from Trolling Data

Sayef Iqbal, Fazel Keshtkar, Soon Ae Chun
Department of Computer Science
St. John's University and City University of New York
{sayef.iqbal16, keshtkaf}@stjohns.edu and soon.chun@csi.cuny.edu

Abstract

Trolling has become one of the dark side of using internet and social media where anyone can demonstrate and promote anger and narcissistic behavior. Trolls are meant to cause discourse between users via advancing personal political and social agenda. The goal of this study is to use state-of-the-art computational linguistic approach, and semantic-sentiment extraction techniques to find patterns for the trolling contents. To do so, we employ word embedding technique to explore patterns in the tweet context. We perform part of speech extraction and analysis, n-grams and word cloud analysis from different tweet categories. Finally, we utilize the SentiStrength approach to explore the sentiment rooted in the semantics of the tweets. Our dataset contains 34,000 tweets. The data are categorized as LeftTroll and RightTroll. We applied different feature extraction techniques to explore the context of these trolls and the findings are promising.

Introduction

As the use of internet and social media are on the rise, people are becoming more dependent on getting their news feed from social media which eventually works as the origin or source that shapes their political views.

These platforms give the opportunity for trolls to provoke rumors, false information, and speculation, and to take advantage of other dishonest information to handle user opinion (Derczynski and Bontcheva 2014a). Trolls have been used in various occasion, i.e., writing fakes and untruth comments, promote anger, and other posts in Twitter and social media platform to promote their goals (Derczynski and Bontcheva 2014b), (Cambria et al. 2010).

According to (Bishop 2014), trolling is: 'the activity of twitting or posting messages via a communications network that are intended to be provocative, offensive or menacing'. Those who post such tweets and posts are known as trolls. As stated by (Hardaker 2010), a real intention(s) of trolls are to cause disruption and trigger or exacerbate conflict for the purpose of their own amusement.

The troll's comments may have a negative psychological impact on their targets and victims and possibly others who participated in the same conversation. Therefore, It is imperative to identify these types of misinformation and perhaps even terminate the conversation before it evolves into something psychologically disruptive for the participants. Monitoring conversations by human watch is costly and a labor-intensive task. It can potentially place a severe burden on the moderators, and it may not be an effective solution for huge conversation and posts such as Twitter. These types of misleading information via trolling urged researchers to develop algorithms to automatically determine malicious comments, which we refer them as trolling attempts. In fact, recently, there have been some studies to automatically identify comments containing cyberbullying (Van Hee et al. 2015), which corresponds to the most severe cases of trolling (Bishop 2014). In this paper, we think it's important to determine trolling efforts, but also identify troll behaviours based on their context and overlapping impact. Moreover, it is vital to investigate sentiment, psychological, and opinion impact on social media users.

It has been proved that one of the most important role of trolls is to manipulate opinions. There have been various evidences based on recent elections in USA and other places that trolls could change users' opinion (Im et al. 2019). There is evidence that Russia's Internet Research Agency attempted to interfere with the 2016 U.S. election by running fake accounts on Twitter, often referred to as "Russian trolls" (Im et al. 2019).

The goal of this paper is to investigate the following: first, we aim to create Trolling-based features that might be useful for research communities and new annotated resource for computational modeling of trolling. Second goal,to apply word embedding to measure and find similarities among linguistic terms according to their probabilities properties in our dataset. Using word embedding can help us to understand a term characterized by its company between RightTrolls and LeftTrolls. Using this approach, we are able to categorize the terms that have influence both recipients of Left and Right trolls. It is hard to identify the instances that belongs to both side of ails, therefore, our model can be useful to help a classifier for training with features taken from model. As a result, we are trying to identify and extract to key information from the tweet contents.

The rest of this paper organized as followings: In Section 2, we explain the related work; Section 3 describe our methodology; in Section 4, we provide the results and experiments; and we conclude the paper in Section 5.

Related Work

In this section we review the related work. Based of our review of previous research, studies in this area belong to trolling, abusive language, aggression detection, politeness, and bullying. Bishop (2014) studied the troll's personality, their motivations, consequences, and on the people activity in psychological aspects of trolls. In anther study done by OSullivan and Flanagin (2003), they focused on flaming/trolls, and hostile and aggressive interactions that affect on users (OSullivan and Flanagin 2003). In recent study by Iqbal and Keshtkar (2019), they investigate how cognitive features have been used in aggression languages in social media. They used LWIC (Tausczik and Pennebaker 2010) to investigate phsycolinguistic features of aggressive posts.

Conforti, Pilehvar, and Collier (2018) proposed another research related to fake news detection. They implemented a four-stage system to verify rumor in fake news. They applied tracking and stance detection. Another study proposed by Mojica de la Vega and Ng (2018), they presented a computational modeling of trolling and they applied categorization of trolling attempts in terms of troll's perspective and troll's responders' perspectives and produced an annotated dataset.

Im et al. (2019) studied the impact of trolls by adversarial individuals and organizations, They found that trolls have a potential to substantially negatively impact society. On the computational side, Mihaylov and Nakov (2016) (Mihaylov and Nakov 2016) address the problem of identifying opinion manipulation trolls, including paid trolls in news community forums. In studies Kumar, Spezzano, and Subrahmanian (2014) and (Kumar et al. 2018) they investigate on troll identification, but also in predictions based on non-linguistic information such as number of votes, dates, number of comments and other meta data features.

In another research related to bullying (Xu et al. 2012), they studied bullying traces. Bullying traces are self-reported events of individuals describing being part of bullying events, but they studied what are the real impact of trolling on analyzing retrospective incidents in real-time conversations. In similar Hardaker (2010) argues that trolling cannot be studied using established politeness research categories.

Based on all above studies, there are two differences between our research and related works. One is that trolling is focused about not only abusive language but also a much larger range of language styles and addresses the intentions and interpretations of the Left/Right trolls that they can play different role. Sometimes, Left trolls plays as a nighties and vs Which goes beyond the linguistic dimension. Therefor, it is hard to identify their role as a left or right. Second, we are interested in trolling attempts and context from LeftTroll and RightTroll.

Methodology

In this section we discuss the details of our methodology, dataset, pre-processing, feature extraction, and findings that came out of this research.

Dataset

The dataset comprised of 34,818 tweets with account category label with RightTroll and LeftTroll. Table 1 illustrates the distribution of the categories of left and right trolls. The tweets were code-mixed, i.e., it contains texts in different languages (written in Russian and Chinese but most are in English). However, for our research, only English tweets are considered. After removing non English tweets, the remaining contains of 34,818 trolling tweets that 17,627 tweets are in LeftTroll category and 17,191 tweets are in RightTroll.

Table 1 shows an example of tweets in left and right categories.

Table 1: Trolling Tweets

Tweet	Category
President Trump In America we dont worship government we worship God	LeftTroll
Liberals are the most hate filled, racist, hypocritical, bigoted, irrational :/, mentally unstable creatures on earth. A danger to the world.	RightTroll

Pre-Processing

Pre-processing is the technique of cleaning and normalization of data which may consist in removing less important tokens, words, or characters in a text such as 'a', 'and', '@' and other unnecessary stop words and lowering capitalized words like 'APPLE'. We also normalize the dataset by applying linguistic reduction through stemming and different form of standardization.

The texts contained several unimportant tokens, for instance, urls, numbers, html tags, and special characters which caused noise in the text for analysis. We cleaned the data first by employing NLTK (Natural language and Text Processing Toolkit) (Bird and Loper 2004) stemmer and stopwords package. Here is an example of transformation of text before and after pre-processing:
before: 'President Trump In America we dont worship government we worship God';
after: 'president trump america worship government worship god'

Feature Extraction

Feature extraction is an accurate and concise reduction process of raw data to some grouped data (Features). In this section we describe the features that extracted from the dataset for further processing. Features are included BoW (Bag of Words), unigrams, PoS (Part-of-Speech), Sentiment strength and we applied WordEmbedding representation along with PCA scores in order to explore the characteristics of these trolling tweets.

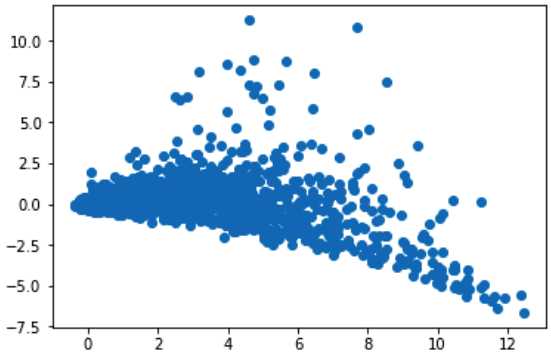

Figure 1: Word Embedding in a scatter plot using both LeftTroll and RightTroll tweets with PCA scores

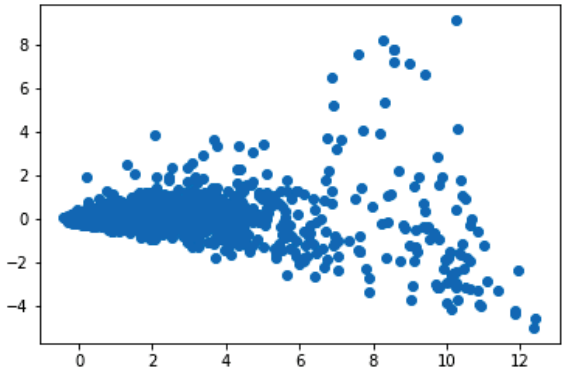

Figure 2: Word Embedding in a scatter plot using LeftTroll tweets with PCA scores

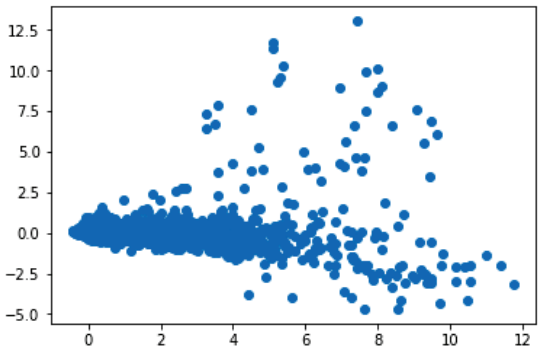

Figure 3: Word Embedding in a scatter plot using RightTroll tweets with PCA scores

Bag of Word Features We extract unigrams, BoGs, from tweets and count their occurrences using bag of word technique. We then use BoGs from each category (LeftTroll and RightTroll) and build a word cloud to visualize most widely used tokens in each of the tweet categories.

Part-of-Speech Features Part-of-Speech (PoS) are classes or lexical representations which have similar grammatical properties. For the purposes of this research, we used NLTK[1] part of speech tagging package to measure the occurrences of PoS tags in each tweet. This led to the extraction of 24 categories of PoSs in our dataset.

SentiStrength Features

We propose a novel approach to compute sentiment and the strength of each word in the dataset and obtained their strength and distribution. We first apply the formula 1 and 2 to calculate polarity of sentiment and the strength of each word.

$$x(w) = r(w).cos(\theta) \quad (1)$$

$$y(w) = r(w).sin(\theta) \quad (2)$$

The r is the weighted *tf-idf* score for each word w and θ are the sentiment polarity score that was calculated using NLTK's vader package. x and y represent the coordinates in the sentiment strength graph for each word w. The approach was adapted from the technique applied by (Saif et al. 2014)

Word Embedding Classification

We used Gensims word2vec [2] package and used our training/test dataset to build word2vec word embedding model. We applied the procedure with LeftTroll data and RightTroll categories and created separate word embedding models. We then represent the word embedding outcomes in scatter plots along with PCA scores and compare contrast and difference word embedding models with words from each category.

As results of Word2vector computation, Figure 1 illustrates the word embedding using both LeftTroll and RightTroll tweets. It also shows the distribution of PCA (Principle Component Analysis) score among the words in the embedding within the same scatter plot.

Figure 2 and 3 represent the word embedding with PCA score as axis for LeftTroll and RightTroll tweets respectively.

Experiments and Results

In this section we discuss some of the key findings from the research.

Here, we present some of the results that we performed on the features that were extracted from the dataset. Table 2 shows the frequency distribution of some of the widely used tokens among the LefTroll tweets.

Word	Frequency
blacklivesmatter	1442
police	769
trump	640
policebrutality	626

Table 2: Examples of words in LeftTroll tweets

On the other hand, Figure 4 represents the word cloud of the tokens (unigram) that were extracted from the LeftTroll

[1] www.nltk.org

[2] https://radimrehurek.com/gensim/models/word2vec.html

Figure 4: Word Cloud for LeftTroll tweets

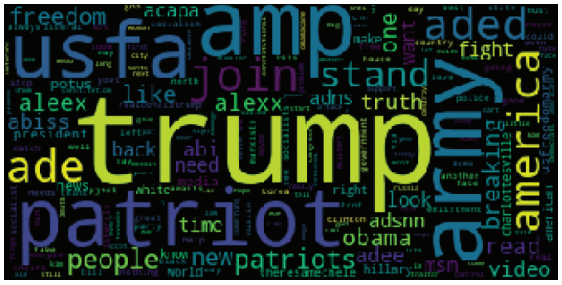

Figure 5: Word Cloud for RightTroll tweets

tweets. And it shown most of the topics are relate to social issues.

Similarly, table 3 shows some of the widely used terms among the RightTroll tweets.

Word	Frequency
trump	2214
army	1984
patriot	1766
america	924

Table 3: Examples of frequent words in RightTroll tweets

Moreover, the word cloud for unigrams extracted from RightTroll tweets is shown in figure 5.

We also extracted the part of speech of each unigram token from both LeftTroll and RightTroll tweets. Figure 6 illustrates the results of PoS and their frequency that extracted among both LeftTroll and RightTroll bag of words, where NN is annotation for noun, NNS for plural noun, JJ and CD for adjective and cardinal numbers respectively.

Figure 7 represents the sentiment scores and the relative strength of each word in the tweets. Our research demonstrates that the tweets have both positive $y > 0$ and negative polarized $y < 0$ tweets but it is worth noting that their strength were mostly 50% effective in expressing the sentiment.

Furthermore, in this research we explored the patterns that obtained in LeftTroll and RightTroll tokens in different word embedding model. These patterns are generated using, i) from both LeftTroll and RightTroll tweet tokens, ii) only LeftTroll tokens from tweets and iii) only RightTroll tokens . from tweets.

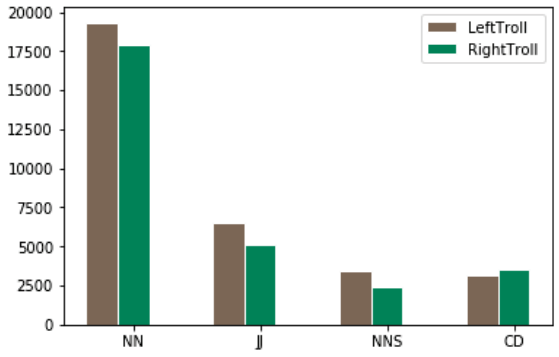

Figure 6: Results of PoS such NN, and JJ in Tweets

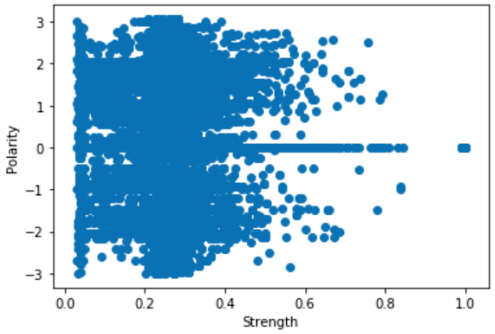

Figure 7: Sentiment-strength of words in tweets

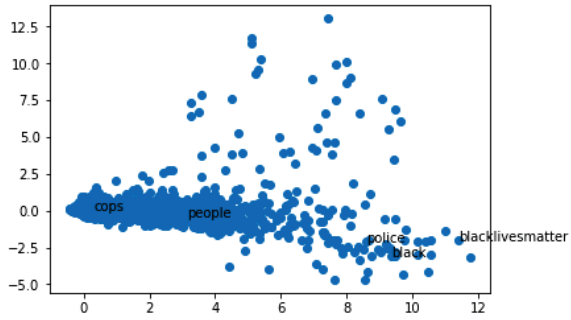

Figure 8: Distribution of frequent LeftTroll tokens in complete word embedding

In another experiment, Figure 8 and 9 demonstrates how patterns of tokens from LeftTroll and RightTroll tweets are distributed in complete word embedding and how their corresponding PCA scores are distributed.

Figure 10 and 11 demonstrates the distribution and PCA scores of frequent tokens from LeftTroll and RightTroll tweets in word embedding generated using LeftTroll tweet tokens respectively.

Similarly, the patterns of the LeftTroll and RightTroll in word embedding generated using RightTroll tweets is demonstrated in figure 12 and 12 respectively.

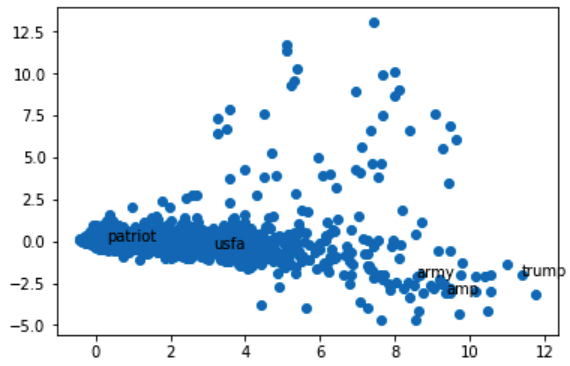

Figure 9: Distribution of frequent RightTroll tokens in complete word embedding

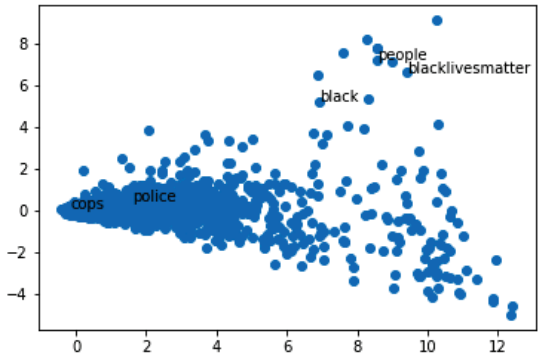

Figure 10: Patterns of LeftTroll tokens in word embedding generated using LeftTroll tokens

Discussion

Most of the LeftTroll tokens were about current social issues including topics such as Black Lives Matter movement and Police Brutality. Also, the LeftTroll tweets tend to explore and discuss issues from the left political spectrum in United States but were often too directed to cops and government for any social issue. Even though the word 'Trump' appears frequently in the LeftTroll tweets however the context can be related to other topics such as police brutality or white supremacy and others. And word cloud generated from tweets in LeftTroll illustrates that. Also, patterns that conducted from word embedding that used both LeftTroll and RightTroll tweets showed Patriot to be the most frequent topic with strong PCA score (lower score) when compared with RightTroll tweets and Cops when compared with LeftTroll tweets. This suggests that there is a strong political division in United Sates even among the trolling social groups, one more vocal about social issues than the other. The findings also suggest the topics like 'black lives matter' appear further away in the word embedding that was developed using RightTroll tweets. On the other hand, tweets from the RightTroll tend to be more about national empowerment, praising current president and at times demeaning minority group.

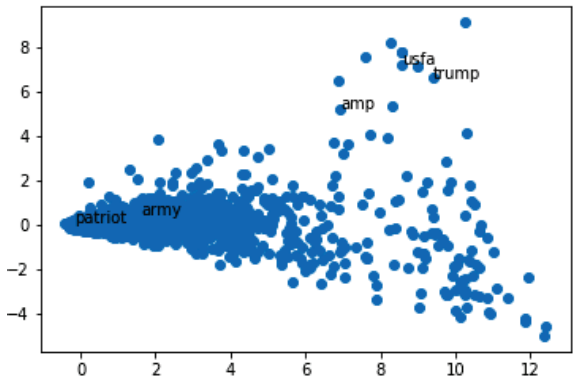

Figure 11: Patterns of RightTroll tokens in word embedding generated using LeftTroll tokens

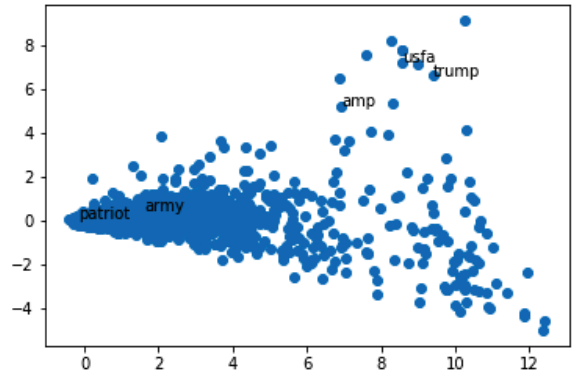

Figure 12: Patterns of LeftTroll tokens in word embedding generated using RightTroll tokens

Conclusion and Future Work

In this paper, we introduce a new computational model based on word embedding and SentiStrength to explore trolling in twitter data. We find semantic patterns that are able to distinguish right and left trolls based on the tweet content. For future work, we aim to extend our analysis on a larger data that include other tweet and trolling categories. We also aim to apply sentiment analysis and opinion mining that trolls can affect on users using SentiStrenth approaches.

References

Bird, S., and Loper, E. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 31. Association for Computational Linguistics.

Bishop, J. 2014. Dealing with internet trolling in political online communities: Towards the this is why we can't have nice things scale. *Int. J. E-Polit.* 5(4):1–20.

Cambria, E.; Chandra, P.; Sharma, A.; and Hussain, A. 2010. Do not feel the trolls. In *In Proceedings of the 3rd International Workshop on Social Data on the Web, SDoW, Shanghai, China*.

Conforti, C.; Pilehvar, M. T.; and Collier, N. 2018. Towards automatic fake news detection: Cross-level stance detection in news articles. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 40–49. Brussels, Belgium: Association for Computational Linguistics.

Derczynski, L., and Bontcheva, K. 2014a. Pheme: Veracity in digital social networks. In *In Proceedings of the UMAP Project Synergy workshop*.

Derczynski, L., and Bontcheva, K. 2014b. Spatio-temporal grounding of claims made on the web, in pheme. In *In Proceedings of the 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, ISA '14, page 65, Reykjavik, Iceland*.

Hardaker, C. 2010. Trolling in asynchronous computermediated communication: From user discussions to academic definitions. In *Journal of Politeness Research, 6(2):215–242*.

Im, J.; Chandrasekharan, E.; Sargent, J.; Lighthammer, P.; Denby, T.; Bhargava, A.; Hemphill, L.; Jurgens, D.; and Gilbert, E. 2019. Still out there: Modeling and identifying russian troll accounts on twitter. *CoRR*.

Iqbal, S., and Keshtkar, F. 2019. Using cognitive learning method to analyze aggression in social media text. In *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019)*. Lecture Notes in Computer Science (LNCS).

Kumar, R.; Reganti, A. N.; Bhatia, A.; and Maheshwari, T. 2018. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.

Kumar, S.; Spezzano, F.; and Subrahmanian, V. S. 2014. Accurately detecting trolls in slashdot zoo via decluttering. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 188–195.

Mihaylov, T., and Nakov, P. 2016. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 399–405. Berlin, Germany: Association for Computational Linguistics.

Mojica de la Vega, L. G., and Ng, V. 2018. Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: European Languages Resources Association (ELRA).

OSullivan, P. B., and Flanagin, A. J. 2003. Reconceptualizing a flaming and other problematic messages. *New Media and Society* 5(1):69–94.

Saif, H.; Fernandez, M.; He, Y.; and Alani, H. 2014. Senticircles for contextual and conceptual semantic sentiment analysis of twitter.

Tausczik, Y., and Pennebaker, J. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. In *Journal of Language and Social Psychology (29)*, 24–54.

Van Hee, C.; Lefever, E.; Verhoeven, B.and Mennes, J.; Desmet, B.; De Pauw, G.; Daelemans, W.; and Hoste, V. 2015. Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*, 672–680.

Xu, J.-M.; Jun, K.-S.; Zhu, X.; and Bellmore, A. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, 656–666. Stroudsburg, PA, USA: Association for Computational Linguistics.