Modeling Trolling in Social Media Conversations

Luis Gerardo Mojica de la Vega and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688, USA
{mojica,vince}@hlt.utdallas.edu

Abstract

Social media websites, electronic newspapers and Internet forums allow visitors to leave comments for others to read and interact. This exchange is not free from participants with malicious intentions, who *troll* others by positing messages that are intended to be provocative, offensive, or menacing. With the goal of facilitating the computational modeling of trolling, we propose a trolling categorization that is novel in the sense that it allows comment-based analysis from both the trolls' and the responders' perspectives, characterizing these two perspectives using four aspects, namely, the troll's intention and his intention disclosure, as well as the responder's interpretation of the troll's intention and her response strategy. Using this categorization, we annotate and release a dataset containing excerpts of Reddit conversations involving suspected trolls and their interactions with other users. Finally, we identify the difficult-to-classify cases in our corpus and suggest potential solutions for them.

Keywords: social media, text categorization, trolling, cyberbullying

1. Introduction

In contrast to traditional content distribution channels like television, radio and newspapers, Internet opened the door for direct interaction between the content creator and its audience. Young people are now gaining more frequent access to online, networked media. Although most of the time, their Internet use is harmless, there are some risks associated with these online activities, such as the use of social networking sites (e.g., Twitter, Facebook, Reddit). The anonymity and freedom provided by social networks makes them vulnerable to threatening situations on the Web, such as *trolling*.

Trolling is "the activity of posting messages via a communications network that are intended to be provocative, offensive or menacing" (Bishop, 2013). People who post such comments are known as *trolls*. According to Hardaker (2010), a troll's "real intention(s) is/are to cause disruption and/or trigger or exacerbate conflict for the purpose of their own amusement". Worse still, the troll's comments may have a negative psychological impact on his target/victim and possibly others who participated in the same conversation. It is therefore imperative to identify such comments and perhaps even terminate the conversation before it evolves into something psychologically disruptive for the participants. Monitoring conversations is a labor-intensive task: it can potentially place a severe burden on the moderators, and it may not be an effective solution when traffic is heavy. This calls for the need to develop automatic methods for identifying malicious comments, which we will refer to as *trolling attempts*.

In fact, there have recently been some attempts to automatically identify comments containing *cyberbullying* (e.g., Van Hee et al. (2015)), which corresponds to the most severe cases of trolling (Bishop, 2013). However, we believe that it is important not only to identify trolling attempts, but also comments that could have a negative psychological impact on their recipients. As an example, consider the situation where a commenter posts a comment with the goal of amusing others. However, it is conceivable that not everybody would be aware of these playful intentions, and these people may disagree or dislike the mocking comments and take them as inappropriate, prompting a negative reaction or psychological impact on themselves.

In light of this discussion, we believe that there is a need to identify not only the trolling attempts, but also comments that could have a negative psychological impact on its receipts. To this end, we seek to achieve the following goals in this paper. First, we propose a comprehensive categorization of trolling that allows us to model not only the troll's intention given his trolling attempt, but also the recipients' *perception* of the troll's intention and subsequently their reaction to the trolling attempt. This categorization gives rise to very interesting problems in pragmatics that involve the computational modeling of intentions, perceived intentions, and reactions to perceived intentions. Second, we create a new annotated resource for computational modeling of trolling. Each instance in this resource corresponds to a *suspected* trolling attempt taken from a Reddit conversation, it's surrounding context, and its immediate responses and will be manually coded with information such as the troll's intention and the recipients' reactions using our proposed categorization of trolling. Finally, we identify the instances that are difficult to classify with the help of a classifier trained with features taken from the state of the art and subsequently present an analysis of these instances.

To our knowledge, our annotated resource is the first one of its sort that allows computational modeling on both the troll's side and the recipients' side. By making it publicly available, we hope to stimulate further research on this task. We believe that it will be valuable to any researcher who is interested in the computational modeling of trolling.

2. Related Work

In this section, we discuss related work in the areas of trolling, bullying, abusive language detection and politeness, as they partially address the problem presented in this

work.
In the realm of psychology, Bishop (2013; 2014) elaborate a deep description of a troll's personality, motivations, effects on the community that trolls interfere in and the criminal and psychological aspects of trolls. Their main focus are flaming (trolls), and hostile and aggressive interactions between users (O'sullivan and Flanagin, 2003).
On the computational side, Mihaylov et al. (2015b; 2015a; 2016) address the problem of identifying opinion manipulation trolls, including paid trolls in news community forums. Not only do they focus solely on troll identification, but the major difference with this work is that all their predictions are based on non-linguistic information such as number of votes, dates, number of comments and so on. In a networks related framework, Kumar et al. (2014) and Guha et al. (2004) present a methodology to identify malicious individuals in a network based solely on the network's properties rather than on the textual content of comments. Cambria et al. (2010) propose a method that involves NLP components, but fail to provide an evaluation of their system.
There is extensive work on detecting offensive and abusive language in social media (Nobata et al., 2016; Xiang et al., 2012). There are two clear differences between their work and ours. One is that trolling is concerned about not only abusive language but also a much larger range of language styles and addresses the intentions and interpretations of the commenters, which goes beyond the linguistic dimension. The other is that we are additionally interested in the reactions to trolling attempts, real or perceived, because we argued that this is a phenomenon that occurs in pairs through the interaction of at least two individuals, which is different from abusive language detection. Also, Xu et al. (2012a; 2012b; 2013) address bullying traces. Bullying traces are self-reported events of individuals describing being part of bullying events, but we believe that the real impact of computational trolling research is not on analyzing retrospective incidents, but on analyzing real-time conversations. Chen et al. (2012) use lexical and semantic features to determine sentence offensiveness levels to identify cyberbullying, offensive or abusive comments on Youtube. On Youtube as well, Dinakar et al. (2012) identified sensitive topics for cyberbullying. Dadvar et al. (2014) used expert systems to classify between bullying and no bullying in posts. Van Hee et al. (2015) predict fine-grained categories for cyberbullying, distinguishing between insults and threats and identified user roles in the exchanges. Finally, Hardaker (2010) argues that trolling cannot be studied using established politeness research categories.

3. Trolling Categorization

In this section, we describe our proposal of a comprehensive trolling categorization. While there have been attempts in the realm of psychology to provide a working definition of trolling (e.g., Hardaker (2010), Bishop (2014)), their focus is mostly on modeling the troll's behavior. For instance, Bishop (2014) constructed a "trolling magnitude" scale focused on the severity of abuse and misuse of internet mediated communications. Bishop (2013) also categorized trolls based on psychological characteristics focused on pathologies and possible criminal behaviors. In contrast, our trolling categorization seeks to model not only the troll's behavior but also the impact on the recipients, as described below.
Since one of our goals is to *identify* trolling events, our datasets will be composed of *suspected* trolling attempts (i.e., comments that are suspected to be trolling attempts). In other words, some of these suspected trolling attempts will be real trolling attempts, and some of them won't. So, if a suspected trolling attempt is in fact not a trolling attempt, then its author will *not* be a troll.
To cover both the troll and the recipients, we define a (suspected trolling attempt, responses) pair as the basic unit that we consider for the study of trolling, where "responses" are all the direct responses to the suspected trolling attempt. We characterize a (suspected trolling attempt, responses) pair using four aspects. Two aspects describe the trolling attempt: (1) **Intention `(I)`** (what is its author's purpose?), and (2) **Intention Disclosure `(D)`** (is its author trying to deceive its readers by hiding his real (i.e., malicious) intentions?). The remaining two aspects are defined on each of the (direct) responses to the trolling attempt: (1) **Intention Interpretation `(R)`** (what is the responder's perception of the troll's intention?), and (2) the **Response strategy `(B)`** (what is the responder's reaction?). Two points deserve mention. First, R can be different from I due to misunderstanding and the fact that the troll may be trying to hide his intention. Second, B is influenced by R, and the responder's comment can itself be a trolling attempt. We believe that these four aspects constitute interesting, under-studied tasks. The possible values of each aspect are described in Table 1.
For a given (suspected trolling attempt, responses) pair, not all of the 189 ($= 3 \times 3 \times 3 \times 7$) combinations of values of the four aspects are possible. There are logical constraints that limit plausible combinations: a) *Trolling* or *Playing Intentions* `(I)` must have *Hidden* or *Exposed* Intention Disclosure `(D)`, b) *Normal* intentions `(I)` can only have *None* Intention disclosure `(D)` and c) *Trolling* or Playing interpretation `(R)` cannot have *Normal* response strategy `(B)`.

3.1. Conversation Excerpts

To enable the reader to better understand this categorization, we present an example excerpt taken from the original (Reddit) conversations. The first comment on the excerpt, generated by author `C0`, is given as a minimal piece of context. The second comment, written by the author `C1` in italics, is the suspected trolling attempt. The rest of the comments comprise all direct responses to the suspected trolling comment.
`C0`: Please post a video of your dog doing this. The way I'm imagining this is adorable.
`C1`: I hope the dog gets run over by a truck on the way out of the childrens playground.
`C2`: If you're going to troll, can you at least try to be a bit more convincing?
`C0`: Haha I hope the cancer kills you.
In this example, we observe that `C0`'s first comment is making a polite request. In return, `C1` made a mean spirited comment whose *intention* is to disrupt and possibly hurt `C0`. Also, `C1`'s comment is not subtle at all, so his inten-

Class	Description	Size %
Intention `(I)`		
Trolling	The comment is malicious in nature, aims to disrupt, annoy, offend, harm or spread purposely false information	53.6% (537)
Mock Trolling or Playing	The comment is playful, joking, teasing or mocking others without the malicious intentions as in the Trolling class	8.9% (89)
No Trolling	A simple comment without malicious or playful intentions.	37.7% (375)
Intention Disclosure `(D)`		
Exposed	A troll, clearly exposing its malicious or playful intentions	34.7% (347)
Hidden	A troll hiding its real malicious or playful intentions	11.5% (115)
None	The comment's author is not a troll, therefore there are no hidden nor exposed malicious or playful intentions	53.8% (539)
Intentions Interpretation `(R)`		
Trolling	The responder believes that the suspected troll is being malicious, annoying, offensive, harmful or attempts to spread false information	59.7% (785)
Mock Trolling or Playing	The responder believes that the suspected troll is being playful, joking, teasing or mocking without the malicious intentions	5.3% (70)
No Trolling	The responder believes that the suspected comment has no malicious intentions nor is playful, it is a simple comment.	35.0% (461)
Response Strategy `(B)`		
Engage	Fall in the perceived provocation, showing an emotional response, upset or annoyed	26.9% (354)
Praise	Acknowledge the perceived malicious or playful intentions and positively recognize the troll's ingenuity or ability	3.0% (39)
Troll	Acknowledge the perceived malicious and counter attack with a trolling attempt	24.0% (316)
Follow	Acknowledge the perceived malicious or playful intention and play along with the troll, further trolling	3.0% (39)
Frustrate	Acknowledges the perceived malicious or playful intentions and attempt to criticize or minimize them	13.0% (171)
Neutralize	Acknowledges the perceived malicious or playful intentions and give no importance to them	9.5% (125)
Normal	There is no perception or interpretation of a trolling attempt and the response is a standard comment	20.7% (272)

Table 1: Classes for trolling aspects: Intention, Intention Disclosure, Intention Interpretation and Response Strategy. *Size* refers to the percentage per class, in parenthesis is the total number of instances in the dataset.

tion is clearly *disclosed*. As for `C2`, she is clearly acknowledging `C1`'s trolling intention and her *response strategy* is *frustrate*. Now, in `C0`'s second comment, we observe that his *interpretation* is clear: he believes that `C1` is *trolling* and the negative effect is so tangible that his response strategy is to *troll* back or *counter-troll* by replying with a comparable mean comment.

4. Corpus and Annotation

Reddit[1] is popular website that allows registered users (without identity verification) to participate in fora grouped by topic or interest. Participation consists of posting stories that can be seen by other users, voting stories and comments, and comments in the story's comment section, in the form of a forum. The forums are arranged in the form of a tree, allowing nested conversations, where the replies to a comment are its direct responses. We collected all comments in the stories' conversation in *Reddit* that were posted in August 2015. Since it is infeasible to manually annotate all of the comments, we process this dataset with the goal of extracting threads that involve suspected trolling attempts and the direct responses to them. To do so, we used Lucene[2] to create an inverted index from the comments. Given that trolling comments have malicious intentions, we queried the inverted index for comments containing the word "**troll**" as well as words having an edit distance of 1 from it. While it is certainly not the case that all trolling comments can be retrieved by this method, we believe such comments would be reasonable candidates of real trolling attempts as our wordlists cover a broad range of strong and weak indicators of malice. This search produced a dataset in which 44.3% of the comments are real trolling attempts. Moreover, it is possible for commenters to believe that they are witnessing a trolling attempt and respond accordingly even where there is none due to misunderstanding. Therefore, the inclusion of comments that do not involve trolling would allow us to learn what triggers a user's interpretation of trolling when it is not present and what kind of response strategies are used.

For each retrieved comment, we reconstructed the original conversation tree it appears in, from the original post (i.e., the root) to the leaves, so that its parent and children can be recovered[3]. We consider a comment in our dataset a suspected trolling attempt if at least one of its immediate children contains the word *troll*. For annotation purposes, we created snippets of conversations exactly like the ones shown in our example, which consists of the parent of the suspected trolling attempt, the suspected trolling attempt, and all of the direct responses to the suspected trolling attempt.

We had two human annotators who were trained on snippets

[1] https://www.reddit.com/

[2] https://lucene.apache.org/

[3] We removed the comments whose text had been deleted

Aspect/Class	mfc F	ngr F	glv F	Size %
I: Intention				
No trolling	69.7	64.2	71.9	53
Trolling	0.0	41.2	27.4	38
Playing	0.0	0.0	0.0	9
Accuracy	53.5	52.8	57.9	-
D: hidden				
Hidden	0.0	0.0	0.0	12
Exposed	0.0	51.2	31.8	35
None	69.9	66.5	72.0	53
Accuracy	53.9	56.8	58.9	-
R: Interpretation				
No trolling	0.0	50.8	38.7	34
Playing	0.0	0.0	0.0	6
Trolling	74.2	74.3	76.3	60
Accuracy	59.2	64.4	64.9	-
B: Response Strategy				
Frustrate	0.0	30.0	19.4	12
Troll	0.0	33.6	27.3	24
Follow	0.0	18.8	6.1	3
Praise	0.0	5.8	0.0	3
Neutralize	0.0	37.0	26.8	9
Normal	0.0	41.2	35.4	20
Engage	43.8	40.5	49.9	29
Accuracy	36.0	36.0	38.0	-
All Tasks Combined				
Total Accuracy	52.5	52.5	54.9	-

Table 2: Results on predicting Intention, Intentions Disclosure, Interpretation, and Response strategy.

(i.e., (suspected trolling attempt, responses) pairs) taken from 200 conversations and were allowed to discuss their findings. After this training stage, we asked them to independently label the four aspects for each snippet. We recognize that this limited amount of information is not always sufficient to recover the four aspects, so we give the annotators the option to discard instances for which they couldn't determine the labels confidently. The final annotated dataset consists of 1000 conversations composed of 6833 sentences and 88047 tokens. The distribution over the classes per trolling aspect is shown in Table 2 in the column "Size".

On the 100 doubly-annotated snippets, we obtained substantial inter-annotator agreement according to Cohen's kappa statistic (Cohen, 1968) for each of the four aspects: Intention: 0.788, Intention Disclosure: 0.780, Interpretation: 0.797 and Response 0.776. In the end, the annotators discussed their discrepancies and managed to resolve all of them.

5. Trolling Attempt Prediction

In this section, we present results on predicting the four aspects of our task.

5.1. Feature Sets

For prediction, we experiment with two feature sets.

N-gram features. We encode each lemmatized and unlemmatized unigram and bigram collected from the training comments as a binary feature. In a similar manner, we include the unigram and bigram along with their POS tag as in Xu et al. (2012a). To extract these features we used Stanford CoreNLP (Manning et al., 2014).

GloVe Embeddings (glv). Word embeddings were created to overcome certain problems with the bag of words (BOW) representation, like sparsity, and weight in correlations of semantically similar words. For this reason, and following Nobota et al. (Nobata et al., 2016), we create a distributed representation of the comments by averaging the word vector of each lowercase token in the comment found in the Twitter corpus pre-trained GloVe vectors (Pennington et al., 2014). The resulting comment vector representation is a 200 dimensional array that is concatenated with the existing BOW.

5.2. Results

Using the features described in the previous subsection, we train four independent classifiers using logistic regression[4], one per each of the four prediction tasks. All the results are obtained using 5-fold cross-validation experiments. In each fold experiment, we use three folds for training, one fold for development, and one fold for testing. All learning parameters are set to their default values except for the regularization parameter, which we tuned on the development set.

Results of the four tasks are shown in Table 2. In each task, we show the F-score on each of its classes in each row as well as the accuracy (the percentage of instances correctly predicted). The last row of the table shows the

[4]We use the *scikit-learn* (Pedregosa et al., 2011) implementation.

total accuracy obtained by averaging the accuracies over the four tasks.

The column mfc shows the results of the frequent class baseline, where the classifier always predicts an instance as belonging to the most frequent class. The percentage of instances belonging to each class can be seen in the last column (Size). As we can see, the mfc baseline achieves an overall accuracy of 52.5.

The next two columns show the results of classifiers trained on exactly one of the two feature groups described in the previous subsection. To get an idea of how strong the mfc baseline is, we can compare it with the classifier trained using only n-gram features (ngr). As we can see, the majority baseline is as strong as ngr w.r.t. overall accuracy (52.5). Nevertheless, ngr is a lot more interesting: it makes predictions on a variety of classes. glv outperforms ngr statistically significantly w.r.t. accuracy on all but Interpretation (paired t-tests, $p < 0.05$). On all but the Response Strategy tasks, glv performs better on the most frequent class than ngr but worse on the second most frequent class.

6. Error Analysis

In order to provide directions for future work, we analyze the errors made by the classifiers on the four prediction tasks.

Errors on *Intention* (I) prediction: The **lack of background** is a major problem when identifying trolling comments. For example, "your comments fit well in Stormfront" seems inoffensive on the surface. However, people who know that Stormfront is a white supremacist website will realize that the author of this comment had an annoying or malicious intention. But our system had no knowledge about it and simply predicted it as non-trolling. These kind of errors reduces recall on the prediction of trolling comments. A solution would be to include additional knowledge from anthologies along with a sentiment or polarity.

Identifying **non-cursing aggressions and insults** is a challenging problem, since the majority of abusive and insulting comments rely on profanity and swearing. The problem arises with subtler aggressions and insults that are equally or even more annoying, such as "Troll? How cute." and "settle down drama queen". The classifier has a more difficult task of determining that these are indeed aggressions or insults. This error also decreases the recall of trolling intention. A solution would be to exploit all the comments made by the suspected troll in the entire conversation in order to increase the chances of finding curse words or other cues that lead the classifier to correctly classify the comment as trolling.

Another source of error is the presence of **controversial topic words** such as "black","feminism", "killing", "racism", "brown", etc. that are commonly used by trolls. The classifier seems too confident to classify a comment as trolling in the presence of these words, but in many cases they do not. In order to ameliorate this problem, one could create ad-hoc word embeddings by training glove or other type of distributed representation on a large corpus for the specific social media platform in consideration. From these vectors one could expect a better representation of controversial topics and their interactions with other words so they might help to reduce these errors.

Errors on *Disclosure* (D) prediction: A major source of error that affects disclosure is the **shallow meaning representation** obtained from the BOW model even when augmented with the distributional features given by the glove vectors. For example, the suspected troll's comment "how to deal with refugees? How about a bullet to the head" is clearly mean-spirited and is an example of disclosed trolling. However, to reach that conclusion the reader need to infer the meaning of "bullet to the head" and that this action is desirable for a vulnerable group like migrants or refugees. This problem produces low recall for the disclosed prediction task. A solution may be the use of deeper semantics, where we represent the comments and sentences in their logical form and infer from them the intended meaning.

Errors on *Interpretation* (R) prediction: it is a common practice from many users to **directly ask** the suspected troll if he/she is trolling or not. There are several variations of this question, such as "Are you a troll?" and "not sure if trolling or not". While the presence of a question like these seems to give us a hint of the responder's interpretation, we cannot be sure of his interpretation without also considering the context. One way to improve interpretation is to exploit the response strategy, but the response strategy in our model is predicted independently of interpretation. So one solution could be similar to the one proposed above for the disclosure task problem: jointly learning classifiers that predict both variables simultaneously.

Errors on *Response Strategy* (B) prediction: In some cases there is a **blurry line between "Frustrate" and "Neutralize"**. The key distinction between them is that there exists some *criticism* in the *Frustrate* responses towards the suspected troll's comment, while "Neutralizing" comments acknowledge that the suspected troll has trolling intentions, but gives no importance to them. For example, response comments such as "oh, you are a troll" and "you are just a lame troll" are examples of this subtle difference. The first is a case of "neutralize" while the second is indeed criticizing the suspected troll's comment and therefore a "frustrate" response strategy. This kind of error affects both precision and recall for these two classes. A possible solution could be to train a specialized classifier to disambiguate between "frustrate" and "neutralize" only.

Another challenging problem is the **distinction between the classes "Troll" and "Engage"**. This is true when the direct responder is intensely flared up with the suspected comment to the point that his own comment becomes a trolling attempt. A useful indicator for distinguishing these cases are the presence of insults, and to detect them we look for swear words, but as we noted before, there is no guarantee that swear words are used for insulting. This kind of error affects the precision and recall for the "troll" and "engage" classes. A solution to this problem may be the inclusion of longer parts of the conversation. It is typical in a troll-engaged comment scheme to observe longer than usual exchanges between two users, and the comments evolve in very agitated remarks. One may then use this information to disambiguate between the two classes.

7. Conclusion

We presented a new view on the computational modeling of trolling in Internet fora where we proposed a comprehensive categorization of trolling attempts that for the first time considers trolling from not only the troll's perspective but also the responders' perspectives. This categorization gives rise to four interesting pragmatics tasks that involve modeling intensions, perceived intensions, and reactions. Perhaps most importantly, we create an annotated dataset that we believe is the first of its sort. We intend to make publicly available with the hope of stimulating research on trolling.

8. Bibliographical References

Bishop, J. (2013). The effect of de-individuation of the internet troller on criminal procedure implementation: An interview with a hater. *International Journal of Cyber Criminology*, 7(1):28.

Bishop, J. (2014). Representations of 'trolls' in mass media communication: a review of media-texts and moral panics relating to 'internet trolling'. *International Journal of Web Based Communities*, 10(1):7–24.

Cambria, E., Chandra, P., Sharma, A., and Hussain, A. (2010). Do not feel the trolls. *ISWC, Shanghai*.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Dadvar, M., Trieschnigg, D., and de Jong, F. (2014). Experts and machines against bullies: a hybrid approach to detect cyberbullies. In *Advances in Artificial Intelligence*, pages 275–281. Springer.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013). A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. (2004). Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412. ACM.

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

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

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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)*, pages 399–405, Berlin, Germany, August. Association for Computational Linguistics.

Mihaylov, T., Georgiev, G. D., Ontotext, A., and Nakov, P. (2015a). Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning, CoNLL*, volume 15, pages 310–314.

Mihaylov, T., Koychev, I., Georgiev, G., and Nakov, P. (2015b). Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 443–450, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Nitta, T., Masui, F., Ptaszynski, M., Kimura, Y., Rzepka, R., and Araki, K. (2013). Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *IJCNLP*, pages 579–586.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Van Hee, C., Lefever, E., Verhoeven, B., 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)*, pages 672–680.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, October.

Xiang, G., Fan, B., Wang, L., Hong, J., and Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceed-*

ings of the 21st ACM international conference on Information and knowledge management, pages 1980–1984. ACM.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012a). 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*, pages 656–666. Association for Computational Linguistics.

Xu, J.-M., Zhu, X., and Bellmore, A. (2012b). Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM.

Xu, J.-M., Burchfiel, B., Zhu, X., and Bellmore, A. (2013). An examination of regret in bullying tweets. In *HLT-NAACL*, pages 697–702.