Hunting for Troll Comments in News Community Forums

Todor Mihaylov
Institute for Computational Linguistics[*]
Heidelberg University
Heidelberg, Germany
mihaylov@cl.uni-heidelberg.de

Preslav Nakov
Qatar Computing Research Institute
Hamad bin Khalifa University
P.O. box 5825, Doha, Qatar
pnakov@qf.org.qa

Abstract

There are different definitions of what a troll is. Certainly, a troll can be somebody who teases people to make them angry, or somebody who offends people, or somebody who wants to dominate any single discussion, or somebody who tries to manipulate people's opinion (sometimes for money), etc. The last definition is the one that dominates the public discourse in Bulgaria and Eastern Europe, and this is our focus in this paper.

In our work, we examine two types of opinion manipulation trolls: paid trolls that have been revealed from leaked "reputation management contracts" and "mentioned trolls" that have been called such by several different people. We show that these definitions are sensible: we build two classifiers that can distinguish a post by such a *paid troll* from one by a *non-troll* with 81-82% accuracy; the same classifier achieves 81-82% accuracy on so called *mentioned troll* vs. *non-troll* posts.

1 Introduction

The practice of using Internet trolls for opinion manipulation has been reality since the rise of Internet and community forums. It has been shown that user opinions about products, companies and politics can be influenced by opinions posted by other online users in online forums and social networks (Dellarocas, 2006). This makes it easy for companies and political parties to gain popularity by paying for "reputation management" to people that write in discussion forums and social networks fake opinions from fake profiles.

Opinion manipulation campaigns are often launched using "personal management software" that allows a user to open multiple accounts and to appear like several different people. Over time, some forum users developed sensitivity about trolls, and started publicly exposing them. Yet, it is hard for forum administrators to block them as trolls try formally not to violate the forum rules. In our work, we examine two types of opinion manipulation trolls: paid trolls that have been revealed from leaked "reputation management contracts"[1] and "mentioned trolls" that have been called such by several different people.

2 Related Work

Troll detection was addressed using analysis of the semantics in posts (Cambria et al., 2010) and domain-adapting sentiment analysis (Seah et al., 2015). There are also studies on general troll behavior (Herring et al., 2002; Buckels et al., 2014).

Astroturfing and misinformation have been addressed in the context of political elections using mapping and classification of massive streams of microblogging data (Ratkiewicz et al., 2011). *Fake profile detection* has been studied in the context of cyber-bullying (Galán-García et al., 2014).

A related research line is on *offensive language* use (Xu and Zhu, 2010). This is related to *cyber-bullying*, which has been detected using sentiment analysis (Xu et al., 2012), graph-based approaches over signed social networks (Ortega et al., 2012; Kumar et al., 2014), and lexico-syntactic features about user's writing style (Chen et al., 2012).

---

[*]This research started in the Sofia University.

[1]The independent Bulgarian media Bivol published a leaked contract described the following services in favor of the government:*"Monthly posting online of 250 comments by virtual users with varied, typical and evolving profiles from different (non-recurring) IP addresses to inform, promote, balance or counteract. The intensity of the provided online presence will be adequately distributed and will correspond to the political situation in the country."* See `https://bivol.bg/en/category/b-files-en/b-files-trolls-en`

Object	Count
Publications	34,514
Comments	1,930,818
-of which replies	897,806
User profiles	14,598
Topics	232
Tags	13,575

Table 1: Statistics about our dataset.

Label	Comments
Paid troll comments	650
Mentioned troll comments	578
Non-troll comments	650+578

Table 2: Comments selected for experiments.

Trustworthiness of statements on the Web is another relevant research direction (Rowe and Butters, 2009). Detecting untruthful and deceptive information has been studied using both psychology and computational linguistics (Ott et al., 2011).

A related problem is *Web spam detection*, which has been addressed using spam keyword spotting (Dave et al., 2003), lexical affinity of arbitrary words to spam content (Hu and Liu, 2004), frequency of punctuation and word co-occurrence (Li et al., 2006). See (Castillo and Davison, 2011) for an overview on adversarial web search.

In our previous work, we focused on finding opinion manipulation troll *users* (Mihaylov et al., 2015a) and on modeling the behavior of exposed vs. paid trolls (Mihaylov et al., 2015b). Here, we go beyond user profile and we try to detect individual troll vs. non-troll *comments* in a news community forum based on both text and metadata.

3 Data

We crawled the largest community forum in Bulgaria, that of Dnevnik.bg, a daily newspaper (in Bulgarian) that requires users to be signed in order to read and comment. The platform allows users to comment on news, to reply to other users' comments and to vote on them with thumbs up/down. We crawled the *Bulgaria*, *Europe*, and *World* categories for the period 01-Jan-2013 to 01-Apr-2015, together with comments and user profiles: 34,514 publications on 232 topics with 13,575 tags and 1,930,818 comments (897,806 of them replies) by 14,598 users; see Table 1. We then extracted comments by *paid trolls* vs. *mentioned trolls* vs. *non-trolls*; see Table 2.

Paid troll comments: We collected them from the leaked reputation management documents, which included 10,150 paid troll comments: 2,000 in Facebook, and 8,150 in news community forums. The latter included 650 posted in the forum of Dnevnik.bg, which we used in our experiments.

Mentioned troll comments: We further collected 1,140 comments that have been replied to with an accusation of being troll comments. We considered a comment as a potential accusation if (*i*) it was a reply to a comment, and (*ii*) it contained words such as *troll* or *murzi(lka)*.[2] Two annotators checked these comments and found 578 actual accusations. The inter-annotator agreement was substantial: Cohen's Kappa of 0.82. Moreover, a simple bag-of-words classifier could find these 578 accusations with an F_1-score of 0.85. Here are some examples (translated):

Accusation: *"To comment from "Prorok Ilia": I can see that you are a red troll by the words that you are using"*

Accused troll's comment: *This Boyko[3] is always in your mind! You only think of him. We like Boko the Potato (the favorite of the Lamb), the way we like the Karlies.*

Paid troll's comment: *in the previous protests, the entire country participated, but now we only see the paid fans of GERB.[4] These are not true protests, but chaotic happenings.*

Non-troll comments are those posted by users that have at least 100 comments in the forum and have never been accused of being trolls. We selected 650 non-troll comments for the paid trolls, and other 578 for the mentioned trolls as follows: for each paid or mentioned troll comment, we selected a non-troll comment at random *from the same thread*. Thus, we have two separate non-troll sets of 650 and of 578 comments.

4 Features

We train a classifier to distinguish *troll* (*paid* or *mentioned*) vs. *non-troll* comments using the following features:

Bag of words. We use words and their frequencies as features, after stopword filtering.[5]

Bag of stems. We further experiment with bag of stems, where we stem the words with the BulStem stemmer (Nakov, 2003a; Nakov, 2003b).

Word n-grams. We also experiment with 2- and 3-word n-grams.

---

[2]Commonly believed in Bulgaria to mean *troll* in Russian (which it does not).

[3]The Bulgarian Prime Minister Mr. Boyko Borisov.

[4]Boyko Borisov's party GERB had fallen down due to protests and here is being accused of organizing protests in turn against the new Socialist government that replaced it.

[5]http://members.unine.ch/jacques.savoy/clef/bulgarianST.txt

Char n-grams. We further use character n-grams, where for each word token we extract all n consecutive characters. We use n-grams of length 3 and 4 only as other values did not help.

Word prefix. For each word token, we extract the first 3 or 4 consecutive characters.

Word suffix. For each word token, we take the last 3 or 4 consecutive characters.

Emoticons. We extract the standard HTML-based emoticons used in the forum of Dnevnik.bg.

Punctuation count. We count the number of exclamation marks, dots, and question marks, both single and elongated, the number of words, and the number of ALL CAPS words.

Metadata. We use the time of comment posting (worktime: 9:00-19:00h vs. night: 21:00-6:00h), part of the week (workdays: Mon-Fri vs. weekend: Sat-Sun), and the rank of the comment divided by the number of comments in the thread.

Word2Vec clusters. We trained word2vec on 80M words from 34,514 publications and 1,930,818 comments in our forum, obtaining 268,617 word vectors, which we grouped into 5,372 clusters using K-Means clustering, and then we use these clusters as features.

Sentiment. We use features derived from MPQA Subjectivity Lexicon (Wilson et al., 2005) and NRC Emotion Lexicon (Mohammad and Turney, 2013) and the lexicon of Hu and Liu (2004). Originally these lexicons were built for English, but we translated them to Bulgarian using Google Translate. Then, we reused the sentiment analysis pipeline from (Velichkov et al., 2014), which we adapted for Bulgarian.

Bad words. We use the number of bad words in the comment as a feature. The words come from the *Bad words list v2.0*, which contains 458 bad words collected for a filter of forum or IRC channels in English.[6] We translated this list to Bulgarian using Google Translate and we removed duplicates to obtain *Bad_Words_Bg_1*. We further used the above word2vec model to find the three most similar words for each bad word in *Bad_Words_Bg_1*, and we constructed another lexicon: *Bad_Words_Bg_3*.[7] Finally, we generate two features: one for each lexicon.

---

[6] http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/

[7] https://github.com/tbmihailov/gate-lang-bulgarian-gazetteers/ - GATE resources for Bulgarian, including sentiment lexicons, bad words lexicons, politicians' names, etc.

Mentions. We noted that trolls use diminutive names or humiliating nicknames when referring to politicians that they do not like, but use full or family names for people that they respect. Based on these observations, we constructed several lexicons with Bulgarian politician names, their variations and nicknames (see footnote 7), and we generated a mention count feature for each lexicon.

POS tag distribution. We also use features based on part of speech (POS). We tag using GATE (Cunningham et al., 2011) with a simplified model trained on a transformed version of the BulTreeBank-DP (Simov et al., 2002). For each POS tag type, we take the number of occurrences in the text divided by the total number of tokens. We use both fine-grained and course-grained POS tags, e.g., from the POS tag *Npmsi*, we generate three tags: *Npmsi*, *N* and *Np*.

Named entities. We also use the occurrence of named entities as features. For extracting named entities such as *location*, *country*, *person_name*, *date_unit*, etc., we use the lexicons that come with Gate's ANNIE (Cunningham et al., 2002) pipeline, which we translated to Bulgarian. In future work, we plan to use a better named entity recognizer based on CRF (Georgiev et al., 2009).

5 Experiments and Evaluation

We train and evaluate an L2-regularized Logistic Regression with LIBLINEAR (Fan et al., 2008) as implemented in SCIKIT-LEARN (Pedregosa et al., 2011), using scaled and normalized features to the [0;1] interval. As we have perfectly balanced sets of 650 positive and 650 negative examples for *paid troll* vs. *non-trolls* and 578 positive and 578 negative examples for *mentioned troll* vs. *non-trolls*, the baseline accuracy is 50%. Below, we report F-score and accuracy with cross-validation.

Table 3, shows the results for experiments to distinguish comments by mentioned trolls vs. such by non-trolls, using all features, as well as when excluding individual feature groups. We can see that excluding character n-grams, word suffixes and word prefixes from the features, as well as excluding bag of words with stems or stop words, yields performance gains; the most sizable gain is when excluding char n-grams, which yields one point of improvement. Excluding bad words usage and emoticons also improves the performance but insignificantly, which might be because they are covered by the bag of words features.

Features	F	Acc
All − char n-grams	79.24	78.54
All − word suff	78.58	78.20
All − word preff	78.51	78.02
All − bow stems	78.32	77.85
All − bow with stop	78.25	77.77
All − bad words	78.10	77.68
All − emoticons	78.08	77.76
All − mentions	78.06	77.68
All	78.06	77.68
All − (bow, no stop)	78.04	77.68
All − NE	77.98	77.59
All − sentiment	77.95	77.51
All − POS	77.80	77.33
All − w2v clusters	77.79	77.25
All − word 3-grams	77.69	77.33
All − word 2-grams	77.62	77.25
All − punct	77.29	76.90
All − metadata	70.77	70.94
Baseline	50.00	50.00

Table 3: **Mentioned troll vs. non-troll comments.** Ablation excluding feature groups.

Features	F	Acc
All − char n-grams	81.08	81.77
All − word suff	81.00	81.77
All − word preff	80.83	81.62
All − bow with stop	80.67	81.54
All − sentiment	80.63	81.46
All − word 2-grams	80.62	81.46
All − w2v clusters	80.54	81.38
All − word 3-grams	80.46	81.38
All − punct	80.40	81.23
All − mentions	80.40	81.31
All	80.40	81.31
All − bow stems	80.37	81.31
All − emoticons	80.33	81.15
All − bad words	80.09	81.00
All − NE	80.00	80.92
All − POS	79.77	80.69
All − (bow, no stop)	79.46	80.38
All − metadata	75.37	76.62
Baseline	50.00	50.00

Table 4: **Paid troll vs. non-troll comments.** Ablation excluding feature groups.

Excluding any of the other features hurts performance, the two most important features to keep being metadata (as it allows us to see the time of posting), and bag of words without stopwords (which looks at the vocabulary choice that mentioned trolls use differently from regular users).

Table 4 shows the results for telling apart comments by paid trolls vs. such by non-trolls, using cross-validation and ablation with the same features as for the mentioned trolls. There are several interesting observations we can make. First, we can see that the overall accuracy for finding paid trolls is slightly higher, namely 81.02, vs. 79.24 for mentioned trolls. The most helpful feature again is metadata, but this time it is less helpful (excluding it yields a drop of 5 points vs. 8 points before). The least helpful feature again are character n-grams. The remaining features fall in between, and most of them yield better performance when excluded, which suggests that there is a lot of redundancy in the features.

Next, we look at individual feature groups. Table 5 shows the results for comments by mentioned trolls vs. such by non-trolls. We can see that the metadata features are by far the most important: using them alone outperforms the results when using all features by 3.5 points.

The reason could be that most troll comments are replies to other comments, while those by non-trolls are mostly not replies. Adding other features such as sentiment-based features, bad words, POS, and punctuation hurts the performance significantly. Features such as bad words are at the very bottom: they do not apply to all comments and thus are of little use alone; similarly for mentions and sentiment features, which are also quite weak in isolation. These results suggest that mentioned trolls are not that different from non-trolls in terms of language use, but have mainly different behavior in terms of replying to other users.

Table 6 shows a bit different picture for comments by paid trolls vs. such by non-trolls. The biggest difference is that metadata features are not so useful. Also, the strongest feature set is the combination of sentiment, bad words distribution, POS, metadata, and punctuation. This suggests that paid trolls are smart to post during time intervals and days of the week as non-trolls, but they use comments with slightly different sentiment and bad word use than non-trolls. Features based on words are also very helpful because paid trolls have to defend pre-specified key points, which limits their vocabulary use, while non-trolls are free to express themselves as they wish.

Features	F	Acc
All	78.06	77.68
Only metadata	84.14	81.14
Sent,bad,pos,NE,meta,punct	77.79	76.73
Only bow, no stop	73.41	73.79
Only bow with stop	73.41	73.44
Only bow stems	72.43	72.49
Only word preff	71.11	71.62
Only w2v clusters	69.85	70.50
Only word suff	69.17	68.95
Only word 2-grams	68.96	69.29
Only char n-grams	68.44	68.94
Only word 3-grams	64.74	67.21
Only POS	64.60	65.31
Sent,bad,pos,NE	63.68	64.10
Only sent,bad	63.66	64.44
Only emoticons	63.30	64.96
Sent,bad,ment,NE	63.11	64.01
Only punct	63.09	64.79
Only sentiment	62.50	63.66
Only NE	62.45	64.27
Only mentions	62.41	64.10
Only bad words	62.27	64.01
Baseline	50.00	50.00

Table 5: **Mentioned troll comments vs. non-troll comments.** Results for individual feature groups.

Features	F	Acc
All	80.40	81.31
Sent,bad,pos,NE,meta,punct	78.04	78.15
Only bow, no stop	75.95	76.46
Only word 2-grams	75.55	74.92
Only bow with stop	75.27	75.62
Only bow stems	75.25	76.08
Only w2v clusters	74.20	74.00
Only word preff	74.01	74.77
Sent,bad,pos,NE	73.89	73.85
Only metadata	73.79	72.54
Only char n-grams	73.02	74.23
Only POS	72.94	72.69
Only word suff	72.03	72.69
Only word 3-grams	69.20	68.00
Only punct	66.80	65.00
Only NE	66.54	64.77
Sent,bad,ment,NE	66.04	64.92
Only sentiment	64.28	62.62
Only mentions	63.28	61.46
Only sent,bad	63.14	61.54
Only emoticons	62.95	61.00
Only bad words	62.22	60.85
Baseline	50.00	50.00

Table 6: **Paid troll vs. non-troll comments.** Results for individual feature groups.

	5	10	15	20
Acc	80.70	81.08	83.41	85.59
Diff	+8.46	+18.51	+30.81	+32.26

Table 7: **Mentioned troll vs. non-troll *users (not comments!)*.** Experiments with different number of minimum mentions for January, 2015. 'Diff" is the difference from the majority class baseline.

6 Discussion

Overall, we have seen that our classifier for telling apart comments by mentioned trolls vs. such by non-trolls performs almost equally well for paid trolls vs. non-trolls, where the non-troll comments are sampled from the same threads that the troll comments come from. Moreover, the most and the least important features ablated from all are also similar. This suggests that mentioned trolls are very similar to paid trolls (except for their reply rate, time and day of posting patterns).

However, using just mentions might be a "witch hunt": some users could have been accused of being "trolls" unfairly. One way to test this is to look not at comments, but at users and to see which users were called trolls by several different other users. Table 7 shows the results for distinguishing users with a given number of alleged troll comments from non-troll users; the classification is based on all comments by the corresponding users. We can see that finding users who have been called trolls more often is easier, which suggests they might be trolls indeed.

7 Conclusion and Future Work

We have presented experiments in predicting whether a comment is written by a troll or not, where we define troll as somebody who was called such by other people. We have shown that this is a useful definition and that comments by mentioned trolls are similar to such by confirmed paid trolls.

References

Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. 2014. Trolls just want to have fun. *Personality and individual Differences*, 67:97–102.

Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. Do not feel the trolls. In *Proceedings of the 3rd International Workshop on Social Data on the Web*, SDoW '10, Shanghai, China.

Carlos Castillo and Brian D. Davison. 2011. Adversarial web search. *Found. Trends Inf. Retr.*, 4(5):377–486, May.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and of the 2012 International Conference on Social Computing*, PASSAT/SocialCom '12, pages 71–80, Amsterdam, Netherlands.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: an architecture for development of robust HLT applications. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 168–175, Philadelphia, Pennsylvania, USA.

Hamish Cunningham, Diana Maynard, and Kalina Bontcheva. 2011. *Text Processing with GATE*. Gateway Press CA.

Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web conference*, WWW '03, pages 519–528, Budapest, Hungary.

Chrysanthos Dellarocas. 2006. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(10):1577–1593.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2014. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. In *Proceedings of the International Joint Conference SOCO13-CISIS13-ICEUTE13*, Advances in Intelligent Systems and Computing, pages 419–428. Springer International Publishing.

Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Petya Osenova, and Kiril Simov. 2009. Feature-rich named entity recognition for Bulgarian using conditional random fields. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '09, pages 113–117, Borovets, Bulgaria.

Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society*, 18(5):371–384.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, Seattle, Washington, USA.

Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. 2014. Accurately detecting trolls in slashdot zoo via decluttering. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, ASONAM '14, pages 188–195, Beijing, China.

Wenbin Li, Ning Zhong, and Chunnian Liu. 2006. Combining multiple email filters based on multivariate statistical analysis. In *Foundations of Intelligent Systems*, pages 729–738. Springer.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015a. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, CoNLL '15, pages 310–314, Beijing, China.

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015b. Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '15, pages 443–450, Hissar, Bulgaria.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Preslav Nakov. 2003a. Building an inflectional stemmer for Bulgarian. In *Proceedings of the 4th International Conference on Computer Systems and Technologies: E-Learning*, CompSysTech '03, pages 419–424, Rousse, Bulgaria.

Preslav Nakov. 2003b. BulStem: Design and evaluation of inflectional stemmer for Bulgarian. In *Proceedings of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics), Thessaloniki, Greece, November, 2003*.

F. Javier Ortega, Jos A. Troyano, Fermn L. Cruz, Carlos G. Vallejo, and Fernando Enrquez. 2012. Propagation of trust and distrust for the detection of trolls in a social network. *Computer Networks*, 56(12):2884 – 2895.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Portland, Oregon.

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

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 249–252, Hyderabad, India.

Matthew Rowe and Jonathan Butters. 2009. Assessing Trust: Contextual Accountability. In *Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web*, SPOT '09, Heraklion, Greece.

Chun-Wei Seah, Hai Leong Chieu, Kian Ming Adam Chai, Loo-Nin Teow, and Lee Wei Yeong. 2015. Troll detection by domain-adapting sentiment analysis. In *Proceedings of the 18th International Conference on Information Fusion*, FUSION '15, pages 792–799, Washington, DC, USA.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Er Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, LREC '02, Canary Islands, Spain.

Boris Velichkov, Borislav Kapukaranov, Ivan Grozev, Jeni Karanesheva, Todor Mihaylov, Yasen Kiprov, Preslav Nakov, Ivan Koychev, and Georgi Georgiev. 2014. SU-FMI: System description for SemEval-2014 task 9 on sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 590–595, Dublin, Ireland.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Vancouver, British Columbia, Canada.

Zhi Xu and Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, CEAS '10, Redmond, Washington, USA.

Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. 2012. Fast learning for sentiment analysis on bullying. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '12, pages 10:1–10:6, Beijing, China.