# Finding Opinion Manipulation Trolls in News Community Forums

**Todor Mihaylov**
FMI
Sofia University
tbmihailov@gmail.com

**Georgi D. Georgiev**
Ontotext AD
Sofia, Bulgaria
georgiev@ontotext.com

**Preslav Nakov**
Qatar Computing Research Institute
HBKU, Qatar
pnakov@qf.org.qa

## Abstract

The emergence of user forums in electronic news media has given rise to the proliferation of *opinion manipulation trolls*. Finding such trolls automatically is a hard task, as there is no easy way to recognize or even to define what they are; this also makes it hard to get training and testing data. We solve this issue pragmatically: we assume that a user who is called a troll by several people is likely to be one. We experiment with different variations of this definition, and in each case we show that we can train a classifier to distinguish a likely troll from a non-troll with very high accuracy, 82–95%, thanks to our rich feature set.

## 1 Introduction

With the rise of social media, it became normal for people to read and follow other users' opinion. This created the opportunity for corporations, governments and others to distribute rumors, misinformation, speculation and to use other dishonest practices to manipulate user opinion (Derczynski and Bontcheva, 2014a). They could consistently use trolls (Cambria et al., 2010), write fake posts and comments in public forums, thus making veracity one of the challenges in digital social networking (Derczynski and Bontcheva, 2014b).

The practice of using opinion manipulation trolls 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 posts by other 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 or companies that write in discussion forums and social networks fake opinions from fake profiles.

In Europe, the problem has emerged in the context of the crisis in Ukraine.[1][2] There have been a number of publications in news media describing the behavior of organized trolls that try to manipulate other users' opinion.[3][4][5] Still, it is hard for forum administrators to block them as trolls try not to violate the forum rules.

## 2 Related Work

Troll detection and offensive language use are understudied problems (Xu and Zhu, 2010). They have been addressed using analysis of the semantics and sentiment in posts to filter out trolls (Cambria et al., 2010); there have been also studies of general troll behavior (Herring et al., 2002; Buckels et al., 2014). Another approach has been to use lexico-syntactic features about user's writing style, structure and specific cyber-bullying content (Chen et al., 2012); cyber-bullying was detected using sentiment analysis (Xu et al., 2012); graph-based approaches over signed social networks have been used as well (Ortega et al., 2012; Kumar et al., 2014). A related problem is that of trustworthiness of statements on the Web (Rowe and Butters, 2009). Yet another related problem is Web spam detection, which has been addressed as a text classification problem (Sebastiani, 2002), e.g., 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.

[1] http://www.forbes.com/sites/peterhimler/2014/05/06/russias-media-trolls/

[2] http://www.theguardian.com/commentisfree/2014/may/04/pro-russia-trolls-ukraine-guardian-online

[3] http://www.washingtonpost.com/news/the-intersect/wp/2014/06/04/hunting-for-paid-russian-trolls-in-the-washington-post-comments-section/

[4] http://www.theguardian.com/world/2015/apr/02/putin-kremlin-inside-russian-troll-house

[5] http://www.theguardian.com/commentisfree/2014/may/04/pro-russia-trolls-ukraine-guardian-online



*Proceedings of the 19th Conference on Computational Language Learning*, pages 310–314,
Beijing, China, July 30-31, 2015. 

| 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.

## 3 Data

We crawled the largest Internet community forum of a Bulgarian media, that of Dnevnik.bg,[6] a daily newspaper that requires users to be signed in in order to comment, which makes it easy to track them. The platform allows users to comment on news, to reply to other users' comments and to vote on them with thumbs up or thumbs down. In the forum, the official language is Bulgarian and all comments are written in Bulgarian.

Each publication has a category, a subcategory, and a list of manually selected tags (keywords). We crawled all publications in the *Bulgaria*, *Europe*, and *World* categories, which turned out to be mostly about politics, for the period 01-Jan-2013 to 01-Apr-2015, together with the comments and the corresponding user profiles as seen in Table 1.

We considered as *trolls* users who were called such by at least $n$ distinct users, and *non-trolls* if they have never been called so. Requiring that a user should have at least 100 comments in order to be interesting for our experiments left us with 317 trolls and 964 non-trolls. Here are two examples (translated):

*"To comment from "Historama": Murzi[7], you know that you cannot manipulate public opinion, right?"*

*"To comment from "Rozalina": You, trolls, are so funny :) I saw the same signature under other comments:)"*

[6] http://dnevnik.bg

[7] *Murzi* is the short for *murzilka*. According to series of recent publications in Bulgarian media, Russian Internet users reportedly use the term *murzilka* to refer to Internet trolls. As a result, this term was adopted by some pro-Western Bulgarian forum users as a way to refer to users that they perceive as pro-Russian opinion manipulation trolls. Despite the term being now in circulation in Bulgaria, it is not really in use in Russia. In fact, the vast majority of Russian Internet users have never heard that *murzilka*, the name of a cute monkey-like children's toy and of a popular Soviet-time children's journal, could possibly be used to refer to Internet trolls.

## 4 Method

Our features are motivated by several publications about troll behavior mentioned above.

For each user, we extract statistics such as number of comments posted, number of days in the forum, number of days with at least one comment, and number of publications commented on. All (other) features are scaled with respect to these statistics, which makes it possible to handle users that registered only recently. Our features can be divided in the following groups:

**Vote-based features.** We calculate the number of comments with positive and negative votes for each user. This is useful as we assume that non-trolls are likely to disagree with trolls, and to give them negative votes. We use the sum from all comments as a feature. We also count separately the comments with high, low and medium positive to negative ratio. Here are some example features: the number of comments where $(\text{positive}/\text{negative}) < 0.25$, and the number of comments where $(\text{positive}/\text{negative}) < 0.50$.

**Comment-to-publication similarity.** These features measure the similarity between comments and publications. We use cosine and TF.IDF-weighted vectors for the comment and for the publication. The idea is that trolls might try to change or blurr the topic of the publication if it differs from his/her views or agenda.

**Comment order-based features.** We count how many user comments the user has among the first $k$. The idea is that trolls might try to be among the first to comment to achieve higher impact.

**Top loved/hated comments.** We calculate the number of times the user's comments were among the top 1, 3, 5, 10 most loved/hated comments in some thread. The idea is that in the comment thread below many publications there are some trolls that oppose all other users, and usually their comments are among the most hated.

**Comment replies-based features.** These are features that count how many user comments are replies to other comments, how many are replies to replies, and so on. The assumption here is that trolls try to post the most comments and want to dominate the conversation, especially when defending a specific cause. We further generate complex features that mix comment reply features and vote counts-based features, thus generating even more features that model the relationship between replies and user agreement/disagreement.

**Time-based features.** We generate features from the number of comments posted during different time periods on a daily or on a weekly basis. We assume that users that write comments on purpose could be paid, or could be activists of political parties, and they probably have some usual times to post, e.g., maybe they do it as a full-time job. On the other hand, most non-trolls work from 9:00 to 18:00, and thus we could expect that they should probably post less comments during this part of the day. We have time-based features that count the number of comments from 9:00 to 9:59, from 12:00 to 12:59, during working hours 9:00-18:00, etc.

All the above features are scaled, i.e., divided by the number of comments, the number of days in the forum, the number of days with more than one comment. Overall, we have a total of 338 such scaled features. In addition, we define a new set of features, which are non-scaled.

**Non-scaled features.** The non-scaled features are features based on the same statistics as above, but just not divided by the number of comments / number of days in the forum / number of days with more that one comment, etc. For example, one non-scaled feature is the number of times a comment by the target user was voted negatively, i.e., as thumbs down, by other users. As a non-scaled feature, we would use this number directly, while above we would scale it by dividing it by the total number of user's comments, by the total number of publications the user has commented on, etc. Obviously, there is a danger in using non-scaled features: older users are likely to have higher values for them compared to recently-registered users.

## 5 Experiments and Evaluation

As we mentioned above, in our experiments, we focus on users with at least 100 comments. This includes 317 trolls and 964 non-trolls. For each user, we extract the above-described features, scaled and non-scaled, and we normalize them in the -1 to 1 interval. We then use a support vector machine (SVM) classifier (Chang and Lin, 2011) with an RBF kernel with C=32 and g=0.0078125 as this was the best-performing configuration. In order to avoid overfitting, we used 5-fold cross-validation. The results are shown in Tables 2 and 3, where the *Accuracy* column shows the cross-validation accuracy and the *Diff* column shows the improvement over the majority class baseline.

| **Features** | **Accuracy** | **Diff** |
|---|---|---|
| AS + Non-scaled | 94.37(+3.74) | 19.13 |
| AS − total comments | 91.17(+0.54) | 15.93 |
| AS − comment order | 91.10(+0.46) | 15.85 |
| AS − similarity | 91.02(+0.39) | 15.77 |
| AS − time day of week | 90.78(+0.15) | 15.53 |
| AS − trigg rep range | 90.78(+0.15) | 15.53 |
| AS − time all | 90.71(+0.07) | 15.46 |
| **All scaled (AS)** | **90.63** | **15.38** |
| AS − top loved/hated | 90.55(-0.07) | 15.30 |
| AS − time hours | 90.47(-0.15) | 15.22 |
| AS − vote u/down rep | 90.47(-0.15) | 15.22 |
| AS − similarity top | 90.32(-0.31) | 15.07 |
| AS − triggered cmnts | 90.32(-0.31) | 15.07 |
| AS − is rep to has rep | 90.08(-0.54) | 14.83 |
| AS − vote up/down all | 89.69(-0.93) | 14.44 |
| AS − is reply | 89.61(-1.01) | 14.36 |
| AS − up/down votes | 88.29(-2.34) | 13.04 |

Table 2: Results for classifying 317 mentioned trolls vs. 964 non-trolls for *All Scaled (AS)* '–' (minus) some scaled feature group. The *Accuracy* column shows the cross-validation accuracy, and the *Diff* column shows the improvement over the majority class baseline.

Table 2 presents the results when using all features, as well as when using all features but excluding/adding one feature group. Here *All scaled (AS)* refers to the features from all groups except for those in the *non-scaled features* group described last in the previous section.

We can see that the best feature set is the one that includes all features, including the *Non-scaled* features group: adding this group contributes +3.74 to accuracy. We further see that excluding features based on time, e.g., *AS − time day of week* and *AS − time all*, improves the accuracy, which means that time of posting is not so important as a feature. Similarly, we see that it hurts accuracy to use as features the total number or the order of comments. Finally, the most important features turn out to be those based on replies and on thumbs up/down votes.

Next, Table 3 shows results of experiments when using different feature groups in isolation. As expected, the features that hurt most when excluded from the *All scaled* feature set, perform best when used alone. Here, the *similarity* features perform worst, which suggests that trolls tend not to change the topic.

| Features | Accuracy | Diff |
|---|---|---|
| All Non-scaled | 93.21 | 17.95 |
| Only vote up/down | 87.67 | 12.41 |
| Only vote up/down totals | 87.20 | 11.94 |
| Only reply up/down voted | 86.10 | 10.85 |
| Only time hours | 84.93 | 9.68 |
| Only time all | 84.31 | 9.06 |
| Only is reply with rep | 82.83 | 7.57 |
| Only triggered rep range | 82.83 | 7.57 |
| Only day of week | 82.28 | 7.03 |
| Only total comments | 82.28 | 7.03 |
| Only reply status | 80.72 | 5.46 |
| Only triggered replies | 80.33 | 5.07 |
| Only comment order | 80.09 | 4.84 |
| Only top loved/hated | 79.39 | 4.14 |
| Only pub similarity top | 75.25 | 0.00 |
| Only pub similarity | 75.25 | 0.00 |

Table 3: Results for classifying 317 mentioned trolls vs. 964 non-trolls for individual feature groups (all scaled, except for line 1). The *Accuracy* column shows the cross-validation accuracy, and the *Diff* column shows the improvement over the majority class baseline.

## 6 Discussion

We considered as trolls people who try to manipulate other users' opinion. Our positive and negative examples are based on trolls having been called such by at least $n$ other users (we used $n = 5$).

However, this is much of a witch hunt and despite our good overall results, the data needs some manual checking in future work. We are also aware that some trolls can actually accuse non-trolls of being trolls, and we cannot be sure whether this is true or not unless we have someone to check it manually. In fact, we do have a list of trolls that are known to have been paid (as exposed in Bulgarian media), but there are only 15 of them, and we could not build a good classifier using only them due to the severe class imbalance.

As the choice of a minimum number of accusations for a user of being a troll that we used to define a troll, namely $n = 5$, might be seen as arbitrary, we also experimented with $n$ = 3, 4, 5, 6, while keeping the required minimum number of comments per user to be 100 as before. The results are shown in Table 4. We can see that as the number of troll mentions/accusations increases, so does the cross-validation accuracy.

| min mentions | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| trolls | 545 | 419 | 317 | 260 |
| non-troll | 964 | 964 | 964 | 964 |
| Accuracy | 85.49 | 87.85 | 90.87 | 92.32 |
| Diff | +21.60 | +18.15 | +15.61 | +13.56 |

Table 4: Results for classifying mentioned trolls vs. non-trolls, using different numbers of minimum troll accusations to define a troll (users with 100 comments or more only). The *Accuracy* column shows the cross-validation accuracy, and the *Diff* column shows the improvement over the majority class baseline.

However, this is partly due to the increased class imbalance of trolls vs. non-trolls, which can be seen by the decrease in the improvement of our classifier compared to the majority class baseline.

We also ran experiments with a fixed number of minimum mentions for the trolls (namely 5 as before), but with varying minimum number of comments per user: 10, 25, 50, 100. The results are shown in Figure 1. We can see that as the minimum number of comments increases, the cross-validation accuracy for both the baseline and for our classifier decreases (as the troll-vs-non-troll ratio becomes more balanced); yet, the improvement of our classifier over the baseline increases, which means that the more we know about a user, the better we can predict whether s/he will be seen as a troll by other users.

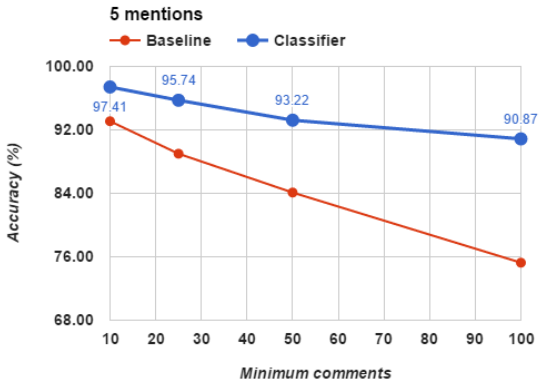


Figure 1: Results for classifying mentioned trolls vs. non-trolls for users with a different minimal number of comments (trolls were accused of being such by 5 or more different users). Shown are results for our classifier and for the majority class baseline.

## 7 Conclusion and Future Work

We have presented experiments in trying to distinguish trolls vs. non-trolls in news community forums. We have experimented with a large number of features, both scaled and non-scaled, and we have achieved very strong overall results using statistics such as number of comments, of positive and negative votes, of posting replies, activity over time, etc. The nature of our features means that our troll detection works best for "elder trolls" with at least 100 comments in the forum. In future work, we plan to add content features such as keywords, topics, named entities, part of speech, and named entities, which should help detect "fresh" trolls. Our ultimate objective is to be able to find and expose *paid* opinion manipulation trolls.

## Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments, which have helped us to improve the paper.

## 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.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

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.

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.

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

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

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 tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, Seattle, WA, 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.

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

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.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

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*.

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, New York, NY, USA.