

# Exposing Paid Opinion Manipulation Trolls

Todor Mihaylov, Ivan Koychev

FMI

Sofia University

tbmihailov@gmail.com

koychev@fmi.uni-sofia.bg

Georgi D. Georgiev

Ontotext AD

Sofia, Bulgaria

georgiev@ontotext.com

Preslav Nakov

Qatar Computing Research Institute

HBKU, Qatar

pnakov@qf.org.qa

## Abstract

Recently, Web forums have been invaded by *opinion manipulation trolls*. Some trolls try to influence the other users driven by their own convictions, while in other cases they can be organized and paid, e.g., by a political party or a PR agency that gives them specific instructions what to write. Finding paid trolls automatically using machine learning is a hard task, as there is no enough training data to train a classifier; yet some test data is possible to obtain, as these trolls are sometimes caught and widely exposed. In this paper, we solve the training data problem by assuming that a user who is called a *troll* by several different people is likely to be such, and one who has never been called a troll is unlikely to be such. We compare the profiles of (i) paid trolls vs. (ii) “mentioned” trolls vs. (iii) non-trolls, and we further show that a classifier trained to distinguish (ii) from (iii) does quite well also at telling apart (i) from (iii).

## 1 Introduction

During the 2013-2014 Bulgarian protests against the Oresharski cabinet, social networks and news community forums became the main “battle grounds” between supporters and opponents of the government. In that period, there was notable censorship in the media, and many people who lived outside the capital did not really know what was actually happening. Moreover, there was a very notable presence of government supporters in Web forums. In series of leaked documents in the independent Bulgarian media Bivol,<sup>1</sup> it was alleged that the ruling Socialist party was paying Internet trolls with EU Parliament money.

<sup>1</sup><https://bivol.bg/en/category/b-files-en/b-files-trolls-en>

The Bivol’s leaked documents revealed for the first time such a practice by a political party despite the problem with opinion manipulation being generally notable across Eastern Europe. The reputation management documents described the following services: “*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 counter-act. The intensity of the provided online presence will be adequately distributed and will correspond to the political situation in the country.*”

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 (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. Yet, over time, forum users developed sensitivity about trolls, and started publicly exposing them.

## 2 Related Work

A popular way to manipulate public opinion in Internet is by making controversial posts on a specific topic that aim to win the argument at any cost, usually accompanied by untruthful and deceptive information. The problem of deceptive opinion spam is studied in (Ott et al., 2011), where the authors integrated work from both psychology and computational linguistics trying to detect fake opinions that were written to sound authentic. Malicious troll users posting misinformation posts have also been studied using graph-based approaches over signed social networks (Ortega et al., 2012; Kumar et al., 2014). A related problem is that of trustworthiness of statements on the Web (Rowe and Butters, 2009).

Troll detection and offensive language use are understudied problems (Xu and Zhu, 2010). They have been addressed using analysis of the semantics and the sentiment in posts (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 cyber-bullying content (Chen et al., 2012); cyber-bullying was detected using user profile and post metadata (Galn-Garca et al., 2014), and sentiment analysis (Xu et al., 2012).

A related problem is that of Web spam detection, usually addressed as text classification (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.

### 3 Data

We crawled the largest media community forum in Bulgaria, that of Dnevnik.bg<sup>2</sup>, a daily newspaper that requires users to be signed in to comment (all in Bulgarian), 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/down. Each publication has a category, a subcategory, and a list of manually selected tags (keywords).

We crawled the *Bulgaria*, *Europe*, and *World* categories for the period 01-Jan-2013 to 01-Apr-2015, together with the comments and the corresponding user profiles: 34,514 publications on 232 topics and with 13,575 tags, 1,930,818 comments (897,806 of them replies), and 14,598 users.

We have three groups of users: known paid trolls (as exposed in Bivol), “mentioned” trolls (called trolls by a certain number of different users), and non-trolls (never called trolls, despite having a high number of posts). Looking at users with at least 150 comments, we have 314 “mentioned” trolls (mentioned by five or more users) vs. 964 non-trolls (vs. some in between); we further have 15 paid trolls from Bivol. Here is an example post with troll accusation (translated):

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

<sup>2</sup><http://dnevnik.bg>

### 4 Method

We train a classifier to distinguish “mentioned” trolls vs. non-trolls; we experiment both with balanced and (natural) imbalanced classes. Then, at test time, we evaluate how well the classifier performs at discriminating paid trolls vs. non-trolls. We use a support vector machine (SVM) classifier (Chang and Lin, 2011) with a radial basis function (RBF) kernel, and features motivated by several publications about troll behavior.

Note that we perform the classification at the user level, i.e., based on user activity history, from which we extract statistics summarizing the user activity. In particular, for each user, we count the number of comments posted, the number of days in the forum, the number of days with at least one comment, and the number of publications commented on. All other features are scaled with respect to these statistics, which makes it possible for us to handle users that registered only recently (which we need to do at test time). Our features can be divided in the following general 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: (a) the number of comments where (positive/negative) < 0.25, and (b) the number of comments where (positive/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 comments by a given user are replies to other users' comments, how many are replies to other replies, and so on. The assumption here is that trolls post not only a large number of comments, but also a large number or replies, as they want to dominate the conversation, especially when defending a specific cause. We further generate complex features that combine user 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 who are paid or who could be activists of political parties 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.

Note that all the above features are scaled, i.e., divided by the number of comments, by the number of days the user has spent in the forum, by the number of days in which the user posted more than one comment, etc. 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 they are not divided by the number of comments / number of days in the forum / number of days with more than 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. Yet, we found unscaled features useful in previous experiments (Mihaylov et al., 2015), so we included them here as well.

## 5 Experiments and Evaluation

In previous work (Mihaylov et al., 2015), we have already experiments with distinguishing "mentioned" trolls vs. non-trolls, achieving accuracy of 88-94%. Here, we are interested in discriminating between *paid* trolls and non-trolls.

Unfortunately, we only know fifteen paid trolls (from the publication in Bivol), which is too little to use for training and testing. Thus, we trained on "mentioned" trolls vs. non-trolls, but we then tested on *paid* trolls vs. non-trolls. We focused on the top four known paid trolls with the highest number of posts, as they had more than 100 comments, which means that we had enough information about them.<sup>3</sup> Thus, for testing we used the four trolls with 100 posts or more, to which we added four non-trolls (i.e., users who have never been called *trolls*). For training, we used 314 "mentioned" troll with 150 posts or more, to which we added 314 non-trolls, also with 150+ posts.

For the experiments, we extracted the features described in the previous section, both scaled and non-scaled, and we normalized them in the -1 to 1 interval. We then trained a support vector machine (SVM) classifier (Chang and Lin, 2011) with a radial basis function (RBF) kernel with  $C=32$  and  $g=0.0078125$ . We chose these values using cross-validation on the training dataset. The testing results are shown in Tables 1 and 2.

Table 1 shows that we can find paid trolls with 100% precision and 75% recall, which is quite good. However, we should be very cautious about any conclusions we draw, as we only had eight testing examples. Yet, let us try to do some analysis. First, note that the best F-score is achieved when using All Scaled features. Moreover, features based on reply status, similarity, up/down votes, number of triggered replies seem to have no impact on the classification performance, as excluding them from the All Scaled features does not affect the results either way. However, excluding time-related features and reply comments vote-based features results in bad score, which means that these features have the most impact on finding paid trolls. Finally, excluding all vote-related features results in zero precision and recall on paid trolls evaluation, which means that these features are key for finding paid trolls.

<sup>3</sup>There were six known paid trolls with more than 40 comments, and the remaining nine known paid trolls from Bivol had less than 40 comments.

Features	Accuracy	Precision	Recall	F-score
All Scaled (AS)	0.88	1.00	0.75	0.86
AS - comment order (Scaled - S)	0.88	1.00	0.75	0.86
AS - is reply (S)	0.88	1.00	0.75	0.86
AS - is reply to has reply (S)	0.88	1.00	0.75	0.86
AS - similarity (S)	0.88	1.00	0.75	0.86
AS - similarity top (S)	0.88	1.00	0.75	0.86
AS - top loved hated (S)	0.88	1.00	0.75	0.86
AS - total comments (S)	0.88	1.00	0.75	0.86
AS - triggered replies range (S)	0.88	1.00	0.75	0.86
AS - triggered replies total (S)	0.88	1.00	0.75	0.86
AS - vote up/down total (S)	0.88	1.00	0.75	0.86
AS - time (S)	0.75	1.00	0.50	0.67
AS - time hours (S)	0.75	1.00	0.50	0.67
AS - vote up/down reply status (S)	0.75	1.00	0.50	0.67
AS - time day of week (S)	0.63	1.00	0.25	0.40
AS + Non Scaled (NS)	0.63	1.00	0.25	0.40
AS - vote up/down all (S)	0.38	0.00	0.00	0.00

Table 1: Results for classifying 4 paid trolls vs. 4 non-trolls for All Scaled (AS) ‘-’ (minus) some scaled feature group. We train on 314 “mentioned” trolls vs. 314 non-trolls. (The bottom features are better, as they yield the highest drop in accuracy and F1 when excluded from All Scaled.)

Features	Accuracy	Precision	Recall	F-score
only day of week (S)	0.88	0.80	1.00	0.89
only reply status (S)	0.75	0.75	0.75	0.75
only time hours (S)	0.75	0.75	0.75	0.75
only top loved hated (S)	0.75	1.00	0.50	0.67
only comment order (S)	0.63	0.67	0.50	0.57
only vote up/down is reply (S)	0.63	0.67	0.50	0.57
only similarity top (S)	0.63	1.00	0.25	0.40
only triggered replies range (S)	0.63	1.00	0.25	0.40
only is reply to has reply (S)	0.50	0.50	0.25	0.33
only similarity (S)	0.50	0.50	0.25	0.33
only time (S)	0.50	0.50	0.25	0.33
only total comments (S)	0.50	0.50	0.25	0.33
only triggered replies total (S)	0.50	0.50	0.25	0.33
only vote up/down all (S)	0.50	0.50	0.25	0.33
only vote up/down total (S)	0.50	0.50	0.25	0.33
All Unscaled	0.50	0.00	0.00	0.00

Table 2: Results for classifying 4 paid trolls vs. 4 non-trolls for individual Scaled (S) feature groups. We train on 314 “mentioned” trolls vs. 314 non-trolls. (The top features are better, as they perform well when used alone.)

Table 2 shows the performance of selected feature groups when used in isolation. We can see that features such as time of posting and votes are among the most important ones; yet, in our previous research, we have found them to be virtually irrelevant for finding “mentioned” trolls vs. non-trolls (Mihaylov et al., 2015).

Table 2 also shows that the best score is achieved by the day of the week feature, which confirms our assumption that paid trolls tend to write on working days. Next come the time-related features, which includes hour-related features and number of comments posted during working hours vs. in the evenings.

## 6 Discussion

Recall that our objective in this work was to identify paid opinion manipulation trolls in Internet forums. Unfortunately, we could not train a classifier to do that directly, as we did not have enough known paid trolls. Thus, we resorted to a simple trick: we considered as trolls those users who were accused of being such by other users. The assumption was that some of these “mentioned” trolls could have actually been paid. However, this is much of a witch hunt and despite our good overall results, the training data is not 100% reliable. For example, some trolls, whether paid or not, could have accused some non-trolls of being trolls, by mistake or on purpose.

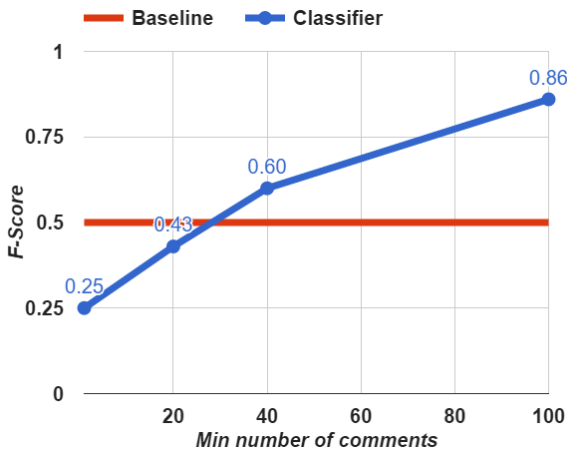


Figure 1: Finding paid trolls with different min number of comments. Training with AS features, and 314 “mentioned” trolls vs. 314 non-trolls.

Recall also that, in our experiments above, we used for testing only four of the fifteen known paid trolls: those with 100 or more comments. It is interesting to see how our classifier would perform if tested on trolls with different minimum number of comments (and the corresponding number of non-trolls). This is shown in Figure 1: we can see that most known paid users with less than 40 comments cannot be exposed as trolls using “mentioned” trolls as training examples.

Next, we vary the number of mentions (by different people) needed for us to consider a user a troll; we try 3, 4 and 6, in addition to 5 as above. Table 3 shows the results when testing on paid trolls with 100+ mentions (4 trolls + 4 non-trolls), where we trained with All Scaled features, and users with 150+ comments and varying minimum number of mentions as a troll.

min mentions	3	4	5	6
“mentioned” trolls	536	416	314	259
non-trolls	536	416	314	259
accuracy	0.75	0.88	0.88	0.75
F-score	0.67	0.86	0.86	0.67

Table 3: Finding paid trolls with 100+ mentions (4 trolls + 4 non-trolls). Training with AS features, and users with 150+ comments and varying minimum number of mentions as a troll.

min mentions	3	4	5	6
“mentioned” trolls	536	416	314	259
non-trolls	536	416	314	259
accuracy	0.83	0.87	0.91	0.92
F-score	0.83	0.87	0.91	0.92

Table 4: Finding “mentioned” trolls (cross-validation on the training dataset). Training with AS features, and users with 150+ comments and varying minimum number of mentions as a troll.

Table 4 shows results when training on the same datasets as in Table 3, but this time evaluating with cross-validation on the training data.

We can see that the best results when testing with paid trolls are achieved for “mentioned” trolls with a minimum of 4 or 5 mentions (Table 3), while when both training and evaluating with “mentioned” trolls (Table 4), the best results are with 6 mentions. This could mean that paid trolls behave more like moderately “mentioned” trolls rather than like highly “mentioned” trolls. More experiments, with a higher number of known paid trolls, are needed in order to confirm this.

Finally, we built and analyzed aggregated profiles for the three kinds of users we considered: (i) paid trolls vs. (ii) “mentioned” trolls vs. (iii) non-trolls.<sup>4</sup> For this purpose, we selected average values for the most notable features for the users with the highest number of comments from each group. We then normalized these values with value/max. The result is shown on Figure 2.

(1 - Active days to all time rate) shows that “mentioned” trolls write at least one comment in 52% of their days of all time being in the forum, while non-trolls do so 36% of the time, and paid trolls only do it 15% of the time. This suggests that paid trolls are less active, maybe because they only write comments when they are paid to do it.

<sup>4</sup>Note that we excluded from our analysis users with too few comments or with too few mentions as a troll.

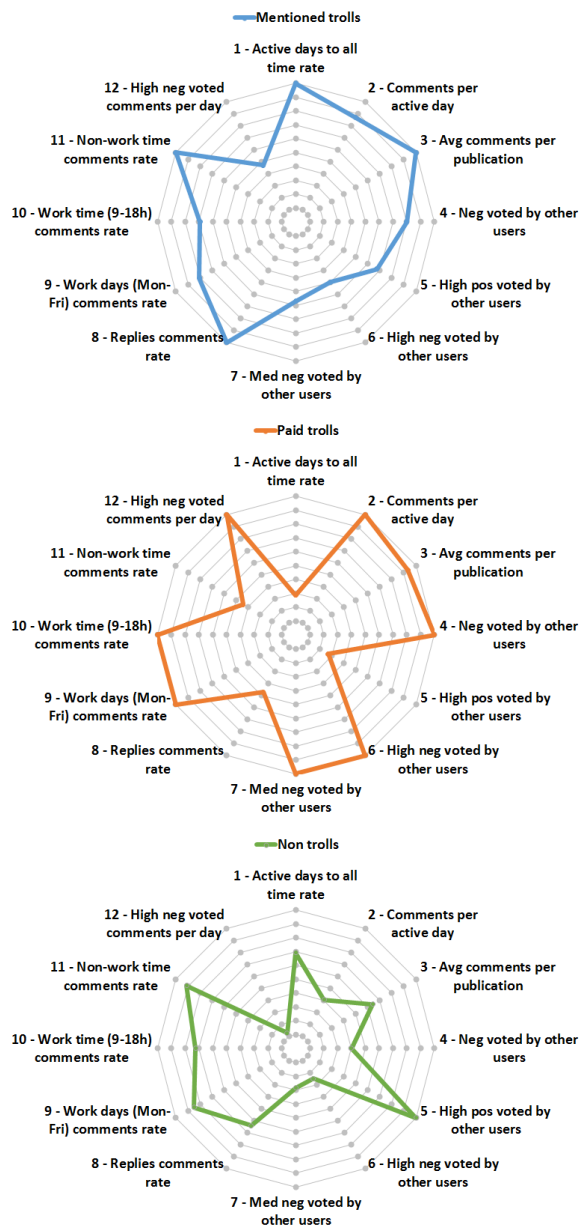


Figure 2: “Mentioned” trolls vs. paid trolls vs. non-trolls based on average feature values.

(2 - Comments per active day) shows that paid trolls and “mentioned” trolls write twice as many comments as non-trolls per day.

(3 - Avg comments per publication) shows that both paid and “mentioned” trolls post more comments per publication than non-trolls.

(4 - Neg voted by other users), (6 - High neg voted by other users), (7 - Med neg voted by other users) show that both paid and “mentioned” trolls have much more negatively voted comments than non-trolls. Yet, this is higher for paid trolls, which could mean that they have more influence compared to the self-driven “mentioned” trolls.

(5) - “mentioned” trolls have more positively voted comments compared to paid trolls.

(8 - Replies comments rate) - “mentioned” trolls are more likely to write comments that are replies to other user’s comments compared to non-trolls, while paid trolls prefer to write specific comments and not to enter personal “battles”. Moreover, paid trolls are more likely to write comments on working days (9 - Work days (Mon-Fri) comments rate) (Mon-Fri), and during working hours (9-18h) ((10 - Work time (9-18h) comments rate),(11 - Non-work time comments rate)) while “mentioned” trolls and non-trolls would write comments at any-time, though mostly during non-working hours.

These observations confirm our assumptions that paid trolls write comments primarily for the money, while “mentioned” trolls do so anytime, and are “self-driven”. Yet, note that some of our “mentioned” trolls might be actually paid.

## 7 Conclusion and Future Work

We have presented experiments in trying to distinguish *paid* opinion manipulation trolls vs. non-trolls in Internet forums. As we did not have enough known paid trolls, for training we used “mentioned” trolls, assuming that a user who is called a *troll* by several different people is likely to be one, while one who has never been called a troll is unlikely to be such. We compared the profiles of (i) paid trolls vs. (ii) “mentioned” trolls vs. (iii) non-trolls, and we have shown that a classifier trained to distinguish (ii) from (iii) does quite well also at telling apart (i) from (iii).

Our further analysis has shown that the most important features were the number of comments, of positive and of negative votes, of posted replies, and the time of commenting. Overall, paid trolls looked roughly like the “mentioned” trolls, except that they were posting most of their comments on working days and during working hours.

Unfortunately, our features only worked well for trolls with high number of posts. Thus, in future work, we plan to add keywords, topics, named entities, sentiment analysis (Kapukaranov and Nakov, 2015; Jovanoski et al., 2015), etc, in order to be able to detect “fresh” trolls; this would require stemming (Nakov, 2003b; Nakov, 2003a), POS tagging (Georgiev et al., 2012), and named entity recognition (Georgiev et al., 2009). We also plan to analyze the comment threads as a whole (Barrón-Cedeño et al., 2015; Joty et al., 2015).

## References

- Alberto Barrón-Cedeño, Simone Filice, Giovanni Da San Martino, Shafiq Joty, Lluís Màrquez, Preslav Nakov, and Alessandro Moschitti. 2015. Thread-level information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL-IJCNLP '15*, pages 687–693, Beijing, China.
- 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. LIB-SVM: 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, PASAT/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.
- Patxi Galn-Garca, JosGaviria de la Puerta, CarlosLaorden Gmez, Igor Santos, and PabloGarca Bringas. 2014. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. In Ivaro Herrero, Bruno Baruque, Fanny Klett, Ajith Abraham, Vclav Snel, Andr C.P.L.F. de Carvalho, Pablo Garca Bringas, Ivan Zelinka, Hector Quintin, and Emilio Corchado, editors, *International Joint Conference SOCO13-CISIS13-ICEUTE13*, volume 239 of *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.
- Georgi Georgiev, Valentin Zhikov, Petya Osenova, Kiril Simov, and Preslav Nakov. 2012. Feature-rich part-of-speech tagging for morphologically complex languages: Application to Bulgarian. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12*, pages 492–502, Avignon, France.
- 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, WA, USA.
- Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global thread-level inference for comment classification in community question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '15*, Lisbon, Portugal.
- Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2015. Sentiment analysis in Twitter for Macedonian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '15*, Hissar, Bulgaria.
- Borislav Kapukaranov and Preslav Nakov. 2015. Fine-grained sentiment analysis for movie reviews in Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP '15*, Hissar, Bulgaria.
- 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. 2015. 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.

- Preslav Nakov. 2003a. Building an inflectional stemmer for Bulgarian. In *Proceedings of the 4th International Conference on Computer Systems and Technologies*, CompSysTech '03, pages 419–424, Sofia, Bulgaria.
- Preslav Nakov. 2003b. BulStem: Design and evaluation of an inflectional stemmer for Bulgarian. In *Proceedings of the Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece.
- 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.
- 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.
- 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*, CEAS '10, pages 20–29, Redmond, WA, USA.
- Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. 2012. Fast learning for sentiment analysis on bullying. In *Proceedings of the International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '12, pages 10:1–10:6, New York, NY, USA.