**2.1.1 FINDING OPINION MANIPULATION TROLLS IN NEWS COMMUNITY FORUMS**

**AUTHORS**

**Todor Mihaylov**

**Georgi D. Georgiev**

**Preslav Nakov**

In this paper [1] the authors stated that 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. They solved this issue: they assumed that a user who is called a troll by several people is likely to be one. They experimented with different variations of this definition, and in each case they showed that they 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.

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. There have been a number of publications in news media describing the behavior of organized trolls that try to manipulate other users’ opinion. Still, it is hard for forum administrators to block them as trolls try not to violate the forum rules.

They crawled the largest Internet community forum of a Bulgarian media, that of Dnevnik.bg, a daily newspaper that requires users to be signed 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.

They 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.

Their 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. They calculated the number of comments with positive and negative votes for each user. This is useful as they assumed that non-trolls are likely to disagree with trolls, and to give them negative votes. They used the sum from all comments as a feature. They also counted separately the comments with high, low and medium  
positive to negative ratio. Here are some example features: the number of comments where  
(positive/negative) < 0.25, and 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. They counted 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. They calculated 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. They further generated 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.

They concluded that they have presented experiments in trying to distinguish trolls vs. non-trolls in news community forums. They have experimented with a large number of features, both scaled and non-scaled, and they 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 their features means  
that our troll detection works best for “elder trolls” with at least 100 comments in the forum. In future work, they plan to add content features such as key words, 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.

**2.1.2 EXPOSING PAID OPINION MANIPULATION TROLLS**

**AUTHORS**

**Todor Mihaylov,**

**Ivan Koychev**

**Georgi D. Georgiev**

**Preslav Nakov**

In this paper [2] the authors stated that 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, they solved 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.

They compared 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). 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, it was alleged that the ruling Socialist party was paying Internet trolls with EU Parliament money. 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 counteract. 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. They concluded that they have presented experiments in trying to distinguish paid opinion manipulation trolls vs. non-trolls in Internet forums. As they 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.

They compared the profiles of (i) paid trolls vs. (ii) “mentioned” trolls vs. (iii) non-trolls, and they have shown that a classifier trained to distinguish (ii) from (iii) does quite well also at telling apart (i) from (iii). Their 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.

**2.1.3 HUNTING FOR TROLL COMMENTS IN NEWS COMMUNITY FORUMS**

**AUTHORS**

**Todor Mihaylov**

**Preslav Nakov**

In this paper [3] the authors stated that 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 their focus in this paper. In their work, they examined 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. They showed that these definitions are sensible: they built 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.

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 their work, they examined 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.

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

**2.1.4 FROM CLICKBAIT TO FAKE NEWS DETECTION: AN APPROACH BASED ON DETECTING THE STANCE OF HEADLINES TO ARTICLES  
AUTHORS**

**Peter Bourgonje,**

**Julian Moreno Schneider,**

**Georg Rehm**

In this paper [4] the authors stated that they presented a system for the detection of  
the stance of headlines with regard to their corresponding article bodies. The approach can be applied in fake news, especially clickbait detection scenarios. The component is part of a larger platform for the curation of digital content; they considered veracity and relevancy an increasingly important part of curating online information. They want to contribute to the  
debate on how to deal with fake news and related online phenomena with technological means, by providing means to separate related from unrelated headlines and further classifying the related headlines.

On a publicly available data set annotated for the stance of headlines with regard to their corresponding article bodies, they achieved a (weighted) accuracy score of 89.59.

With the advent of social media and its increasingly important role as a provider and amplifier of news, basically anyone, anywhere, can produce and help circulate content for other people to read. Traditional barriers to publishing content (like a press to print newspapers or broadcasting time for radio or television) have disappeared, and with this, at least part of traditional quality control procedures had disappeared as well.

Basic journalistic principles like source verification, fact checking and accountability can be easily bypassed or simply ignored by individuals or organisations publishing content on Twitter, Facebook or other social networks.

The impact of this situation is illustrated by the predominance of terms like “trolls”, “fake news”, “post-truth media” and “alternative facts”. There is evidence that these developments and their effects are not harmless but can have a significant impact on real-world events, which is illustrated by a description of the role of social media in the 2016 US presidential election by (Allcott and Gentzkow, 2017), and by a study on the effectiveness and debunking strategies of  
rumours surrounding the Affordable Care Act by (Berinsky, 2017).

While the cause of this situation may have its roots in many different aspects of modern society, and hence needs to be approached from several different angles, they aimed to make a contribution from the angle of Language Technology and Natural Language Processing. They considered fully automated procedures for fact-checking, clickbait detection or fake news classification not feasible at this point (Rehm, 2017), but aim to support the community by providing means of detecting articles or pieces of news that need to be approached  
with caution, where a human has to make final decisions (on credibility, legitimacy etc.), but is  
aided by a set of tools.

The approach described in this paper can serve as the back-end of such a smart set of tooling around fact-checking and can augment news coming from both traditional and non-traditional (social media) sources. They envisioned the resulting set of tools as a collection of expert tools for specific job profiles (like a journalist or a news editor), or in the shape of a simple browser plug-in, flagging unverified or dubious content to the end user.

The work presented in this paper was carried out under the umbrella of a two-year research and technology transfer project, in which a research centre collaborates with four SME partners that face the challenge of having to process, analyse and make sense of large amounts of digital content. The companies cover four different use cases and sectors (Rehm and Sasaki, 2015) including journalism. For these partners they developed a platform that provides access to language and knowledge technologies (Bourgonje et al., 2016a,b).

The services are integrated by the SME partners into their own in-house systems or those of clients. In this paper, we aim to contribute to a first step in battling fake news, often referred to as stance detection, where the challenge is to detect the stance of a claim with regard to another piece of content. Their experiments are based on the setup of the first Fake News Challenge (FNC1).1. In FNC1, the claim comes in the form of a headline, and the other piece of content is an article body. This step may seem, and, in fact, is, a long way from automatically checking the veracity of a piece of content with regard to some kind of ground truth. But the problem lies exactly in the definition of the truth, and the fact that it is sensitive to bias.

Additionally, and partly because of this, annotated corpora, allowing training and  
experimental evaluation, are hard to come by and also often (in the case of fact checker archives) not freely available. We argue that detecting whether a piece of content is related or not related to another piece of content (e. g., headline vs. article body) is an important first step, which would perhaps best be described as clickbait detection (i. e., a headline not related to the actual article is more likely to be clickbait).

Following the FNC1 setup, the further classification of related pieces of content into more fine-grained classes provides valuable information once the “truth” (in the form of a collection of facts) has been established, so that particular pieces of content can be classified as “fake” or, rather, “false”. Since this definitive, resolving collection of facts is usually hard to come by, the challenge of stance detection can be put to use combining the outcome with credibility or reputation scores of news outlets, where several high-credibility outlets disagreeing with a particular piece of content point towards a false claim.

Stance detection can also prove relevant for detecting political bias: if authors on the same end of the political spectrum are more likely to agree with each other, the (political) preference of one author can be induced once the preference of the other author is known. Additionally, the stances of utterances towards a specific piece of content can provide hints on its veracity. (Mendoza et al., 2010) show that the propagation of tweets regarding crisis situations (like natural disasters) differs based on their content: tweets spreading news are affirmed by related tweets, whereas tweets spreading rumors are mostly questioned or denied. In this paper they proposed a solution that involves the human-in-the-loop.

They concluded that they presented a system for stance detection of headlines with regard to their corresponding article bodies. Their system is based on simple, lemmatization based n-gram matching for the binary classification of “related” vs. “unrelated” headline/article pairs. The best results were obtained using a setup where the more fine-grained classification of the “related” pairs (into “agree”, “disagree”, “discuss”) is carried out using a Logistic Regression classifier at first, then three binary classifiers with slightly different training procedures for the cases where the first classifier lacked confidence (i. e., the difference between the best and second-best scoring class was below a threshold). They improved on the accuracy base line set by the organizers of the FNC1 by over 10 points and scored 9th place (out of 50 participants) in the actual challenge.