Predicting the Role of Political Trolls in Social Media

Atanas Atanasov
Sofia University
Bulgaria
amitkov@uni-sofia.bg

Gianmarco De Francisci Morales
ISI Foundation
Italy
gdfm@acm.org

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

Abstract

We investigate the political roles of "Internet trolls" in social media. Political trolls, such as the ones linked to the Russian Internet Research Agency (IRA), have recently gained enormous attention for their ability to sway public opinion and even influence elections. Analysis of the online traces of trolls has shown different behavioral patterns, which target different slices of the population. However, this analysis is manual and labor-intensive, thus making it impractical as a first-response tool for newly-discovered troll farms. In this paper, we show how to automate this analysis by using machine learning in a realistic setting. In particular, we show how to classify trolls according to their political role —left, news feed, right— by using features extracted from social media, i.e., Twitter, in two scenarios: (*i*) in a traditional supervised learning scenario, where labels for trolls are available, and (*ii*) in a distant supervision scenario, where labels for trolls are not available, and we rely on more-commonly-available labels for news outlets mentioned by the trolls. Technically, we leverage the community structure and the text of the messages in the online social network of trolls represented as a graph, from which we extract several types of learned representations, i.e., embeddings, for the trolls. Experiments on the "IRA Russian Troll" dataset show that our methodology improves over the state-of-the-art in the first scenario, while providing a compelling case for the second scenario, which has not been explored in the literature thus far.

1 Introduction

Internet "trolls" are users of an online community who quarrel and upset people, seeking to sow discord by posting inflammatory content. More recently, organized "troll farms" of political opinion manipulation trolls have also emerged. Such farms usually consist of state-sponsored agents who control a set of pseudonymous user accounts and personas, the so-called "sockpuppets", which disseminate misinformation and propaganda in order to sway opinions, destabilize the society, and even influence elections (Linvill and Warren, 2018).

The behavior of political trolls has been analyzed in different recent circumstances, such as the 2016 US Presidential Elections and the Brexit referendum in UK (Linvill and Warren, 2018; Llewellyn et al., 2018). However, this kind of analysis requires painstaking and time-consuming manual labor to sift through the data and to categorize the trolls according to their actions. Our goal in the current paper is to automate this process with the help of machine learning (ML). In particular, we focus on the case of the 2016 US Presidential Elections, for which a public dataset from Twitter is available. For this case, we consider only accounts that post content in English, and we wish to divide the trolls into some of the functional categories identified by Linvill and Warren (2018): *left troll*, *right troll*, and *news feed*.

We consider two possible scenarios. The first, prototypical ML scenario is supervised learning, where we want to learn a function from users to categories $\{$*left*, *right*, *news feed*$\}$, and the ground truth labels for the troll users are available. This scenario has been considered previously in the literature by Kim et al. (2019). Unfortunately, a solution for such a scenario is not directly applicable to a real-world use case. Suppose a new troll farm trying to sway the upcoming European or US elections has just been discovered. While the identities of the accounts might be available, the labels to learn from would not be present. Thus, any supervised machine learning approach would fall short of being a fully automated solution to our initial problem.

A more realistic scenario assumes that labels for troll accounts are *not available*. In this case, we need to use some external information in order to learn a labeling function. Indeed, we leverage more persistent entities and their labels: news media. We assume a learning scenario with distant supervision where labels for news media are available. By combining these labels with a citation graph from the troll accounts to news media, we can infer the final labeling on the accounts themselves without any need for manual labeling.

One advantage of using distant supervision is that we can get insights about the behavior of a newly-discovered troll farm quickly and effortlessly. Differently from troll accounts in social media, which usually have a high churn rate, news media accounts in social media are quite stable. Therefore, the latter can be used as an anchor point to understand the behavior of trolls, for which data may not be available.

We rely on embeddings extracted from social media. In particular, we use a combination of embeddings built on the user-to-user mention graph, the user-to-hashtag mention graph, and the text of the tweets of the troll accounts. We further explore several possible approaches using label propagation for the distant supervision scenario.

As a result of our approach, we improve the classification accuracy by more than 5 percentage points for the supervised learning scenario. The distant supervision scenario has not previously been considered in the literature, and is one of the main contributions of the paper. We show that even by hiding the labels from the ML algorithm, we can recover 78.5% of the correct labels.

The contributions of this paper can be summarized as follows:

- We predict the political role of Internet trolls (*left*, *news feed*, *right*) in a realistic, unsupervised scenario, where labels for the trolls are not available, and which has not been explored in the literature before.
- We propose a novel distant supervision approach for this scenario, based on graph embeddings, BERT, and label propagation, which projects the more-commonly-available labels for news media onto the trolls who cited these media.
- We improve over the state of the art in the traditional, fully supervised setting, where training labels are available.

2 Related Work

2.1 Trolls and Opinion Manipulation

The promise of social media to democratize content creation (Kaplan and Haenlein, 2010) has been accompanied by many malicious attempts to spread misleading information over this new medium, which quickly got populated by *sockpuppets* (Kumar et al., 2017), *Internet water army* (Chen et al., 2013), *astroturfers* (Ratkiewicz et al., 2011), and *seminar users* (Darwish et al., 2017). Several studies have shown that trust is an important factor in online relationships (Ho et al., 2012; Ku, 2012; Hsu et al., 2014; Elbeltagi and Agag, 2016; Ha et al., 2016), but building trust is a long-term process and our understanding of it is still in its infancy (Salo and Karjaluoto, 2007). This makes it easy for politicians and companies to manipulate user opinions in community forums (Dellarocas, 2006; Li et al., 2016; Zhuang et al., 2018).

Trolls. Social media have seen the proliferation of fake news and clickbait (Hardalov et al., 2016; Karadzhov et al., 2017a), aggressiveness (Moore et al., 2012), and trolling (Cole, 2015). The latter often is understood to concern malicious online behavior that is intended to disrupt interactions, to aggravate interacting partners, and to lure them into fruitless argumentation in order to disrupt online interactions and communication (Chen et al., 2013). Here we are interested in studying not just any trolls, but those that engage in opinion manipulation (Mihaylov et al., 2015a,b, 2018). This latter definition of *troll* has also become prominent in the general public discourse recently. Del Vicario et al. (2016) have also suggested that the spreading of misinformation online is fostered by the presence of polarization and echo chambers in social media (Garimella et al., 2016, 2017, 2018).

Trolling behavior is present and has been studied in all kinds of online media: online magazines (Binns, 2012), social networking sites (Cole, 2015), online computer games (Thacker and Griffiths, 2012), online encyclopedia (Shachaf and Hara, 2010), and online newspapers (Ruiz et al., 2011), among others.

Troll detection has been addressed by using domain-adapted sentiment analysis (Seah et al., 2015), various lexico-syntactic features about user writing style and structure (Chen et al., 2012; Mihaylov and Nakov, 2016), and graph-based approaches over signed social networks (Kumar et al., 2014).

Sockpuppet is a related notion, and refers to a person who assumes a false identity in an Internet community and then speaks to or about themselves while pretending to be another person. The term has also been used to refer to opinion manipulation, e.g., in Wikipedia (Solorio et al., 2014). Sockpuppets have been identified by using authorship-identification techniques and link analysis (Bu et al., 2013). It has been also shown that sockpuppets differ from ordinary users in their posting behavior, linguistic traits, and social network structure (Kumar et al., 2017).

Internet Water Army is a literal translation of the Chinese term *wangluo shuijun*, which is a metaphor for a large number of people who are well organized to flood the Internet with purposeful comments and articles. Internet water army has been allegedly used in China by the government (also known as *50 Cent Party*) as well as by a number of private organizations.

Astroturfing is an effort to simulate a political grass-roots movement. It has attracted strong interest from political science, and research on it has focused on massive streams of microblogging data (Ratkiewicz et al., 2011).

Identification of malicious accounts in social media includes detecting spam accounts (Almaatouq et al., 2016; McCord and Chuah, 2011), fake accounts (Fire et al., 2014; Cresci et al., 2015), compromised and phishing accounts (Adewole et al., 2017). Fake profile detection has also been studied in the context of cyber-bullying (Galán-García et al., 2016). A related problem is that of *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).

Trustworthiness and veracity analytics of online statements is an emerging research direction, especially given the recent interest in fake news (Lazer et al., 2018). It is related to trolls, as they often engage in opinion manipulation and rumor spreading (Vosoughi et al., 2018). Research topics include predicting the credibility of information in social media (Ma et al., 2016; Mitra et al., 2017; Karadzhov et al., 2017b; Popat et al., 2017) and political debates (Hassan et al., 2015; Gencheva et al., 2017; Jaradat et al., 2018), as well as stance classification (Mohtarami et al., 2018). For example, Castillo et al. (2011) leverage user reputation, author writing style, and various time-based features, Canini et al. (2011) analyze the interaction of content and social network structure, and Morris et al. (2012) studied how Twitter users judge truthfulness. Zubiaga et al. (2016) study how people handle rumors in social media, and found that users with higher reputation are more trusted, and thus can spread rumors easily. Lukasik et al. (2015) use temporal patterns to detect rumors and to predict their frequency, and Zubiaga et al. (2016) focus on conversational threads. More recent work has focused on the credibility and the factuality in community forums (Nakov et al., 2017; Mihaylova et al., 2018, 2019; Mihaylov et al., 2018).

2.2 Understanding the Role of Political Trolls

None of the above work has focused on understanding the role of political trolls. The only closely relevant work is that of Kim et al. (2019), who predict the roles of the Russian trolls on Twitter by leveraging social theory and Actor-Network Theory approaches. They characterize trolls using the digital traces they leave behind, which is modeled using a time-sensitive semantic edit distance. For this purpose, they use the "IRA Russian Troll" dataset (Linvill and Warren, 2018), which we also use in our experiments. However, we have a very different approach based on graph embeddings, which we show to be superior to their method in the supervised setup. We further experiment with a new, and arguably more realistic, setup based on distant supervision, where labels are not available. To the best of our knowledge, this setup has not been explored in previous work.

2.3 Graph Embeddings

Graph embeddings are machine learning techniques to model and capture key features from a graph automatically. They can be trained either in a supervised or in an unsupervised manner (Cai et al., 2018). The produced embeddings are latent vector representations that map each vertex V in a graph G to a d-dimensional vector. The vectors capture the underlying structure of the graph by putting "similar" vertices close together in the vector space. By expressing our data as a graph structure, we can leverage and extract critical insights about the topology and the contextual relationships between the vertices in the graph.

In mathematical terms, graph embeddings can be expressed as a function $f : V \rightarrow R^d$ from the set of vertices V to a set of embeddings, where d is the dimensionality of the embeddings. The function f can be represented as a matrix of dimensions $|V| \times d$. In our experiments, we train Graph Embeddings in an unsupervised manner by using node2vec (Grover and Leskovec, 2016), which is based on random walks over the graph. Essentially, this is an application of the well-known skip-gram model (Mikolov et al., 2013) from word2vec to random walks on graphs.

Besides *node2vec*, there have been a number of competing proposals for building graph embeddings; see (Cai et al., 2018) for an extensive overview of the topic. For example, *SNE* (Liao et al., 2018) model both the graph structure and some node attributes. Similarly, *Line* (Tang et al., 2015) represent each node as the concatenation of two embedded vectors that model first- and second-order proximity. *TriDNR* (Pan et al., 2016) represents nodes by coupling several neural network models. For our experiments, we use node2vec, as we do not have access to user attributes: the users have been banned from Twitter, their accounts were suspended, and we only have access to their tweets thanks to the "IRA Russian Trolls" dataset.

3 Method

Given a set of known political troll users (each user being represented as a collection of their tweets), we aim to detect their role: *left*, *right*, or *news feed*. Linvill and Warren (2018) describe these roles as follows:

Right Trolls spread nativist and right-leaning populist messages. Such trolls support the candidacy and Presidency of Donald Trump and denigrate the Democratic Party; moreover, they often send divisive messages about mainstream and moderate Republicans.

Left Trolls send socially liberal messages and discuss gender, sexual, religious, and -especially- racial identity. Many tweets are seemed intentionally divisive, attacking mainstream Democratic politicians, particularly Hillary Clinton, while supporting Bernie Sanders prior to the elections.

News Feed Trolls overwhelmingly present themselves as US local news aggregators, linking to legitimate regional news sources and tweeting about issues of local interest.

Technically, we leverage the community structure and the text of the messages in the social network of political trolls represented as a graph, from which we learn and extract several types of vector representations, i.e., troll user embeddings. Then, armed with these representations, we tackle the following tasks:

T1 A fully supervised learning task, where we have labeled training data with example troll and their roles;

T2 A distant supervision learning task, in which labels for the troll roles are *not* available at training time, and thus we use labels for news media as a proxy, from which we infer labels for the troll users.

3.1 Embeddings

We use two graph-based (user-to-hashtag and user-to-mentioned-user) and one text-based (BERT) embedding representations.

3.1.1 U2H

We build a bipartite, undirected User-to-Hashtag (U2H) graph, where nodes are users and hashtags, and there is an edge (u, h) between a user node u and a hashtag node h if user u uses hashtag h in their tweets. This graph is bipartite as there are no edges connecting two user nodes or two hashtag nodes. We run node2vec (Grover and Leskovec, 2016) on this graph, and we extract the embeddings for the users (we ignore the hashtag embeddings). We use 128 dimensions for the output embeddings. These embeddings capture how similar troll users are based on their usage of hashtags.

3.1.2 U2M

We build an undirected User-to-Mentioned-User (U2M) graph, where the nodes are users, and there is an edge (u, v) between two nodes if user u mentions user v in their tweets (i.e., u has authored a tweet that contains "$@v$"). We run node2vec on this graph and we extract the embeddings for the users. As we are interested only in the troll users, we ignore the embeddings of users who are only mentioned by other trolls. We use 128 dimensions for the output embeddings. The embeddings extracted from this graph capture how similar troll users are according to the targets of their discussions on the social network.

3.1.3 BERT

BERT offers state-of-the-art text embeddings based on the Transformer architecture (Devlin et al., 2019). We use the pre-trained BERT-large, uncased model, which has 24-layers, 1024-hidden, 16-heads, and 340M parameters, which yields output embeddings with 768 dimensions. Given a tweet, we generate an embedding for it by averaging the representations of the BERT tokens from the penultimate layer of the neural network. To obtain a representation for a user, we average the embeddings of all their tweets. The embeddings extracted from the text capture how similar users are according to their use of language.

3.2 Fully Supervised Learning (T1)

Given a set of troll users for which we have labels, we use the above embeddings as a representation to train a classifier. We use an L2-regularized logistic regression (LR) classifier. Each troll user is an example, and the label for the user is available for training thanks to manual labeling. We can therefore use cross-validation to evaluate the predictive performance of the model, and thus the predictive power of the features.

We experiment with two ways of combining features: *embedding concatenation* and *model ensembling*. Embedding concatenation concatenates the feature vectors from different embeddings into a longer feature vector, which we then use to train the LR model. Model ensembling instead trains a separate model with each kind of embedding, and then merges the prediction of the different models by averaging the posterior probabilities for the different classes. Henceforth, we denote embedding concatenation with the symbol $\|$ and model ensembling with $\oplus$. For example, U2H $\|$ U2M is a model trained on the concatenation of U2H and U2M embeddings, while U2H $\oplus$ BERT represents the average predictions of two models, one trained on U2H embeddings and one on BERT.

3.3 Distant Supervision (T2)

In the distant supervision scenario, we assume not to have access to user labels. Given a set of troll users without labels, we use the embeddings described in Section 3.1 together with mentions of *news media* by the troll users to create proxy models. We assume that labels for news media are readily available, as they are stable sources of information that have a low churn rate.

We propagate labels from the given media to the troll user that mentions them according to the following media-to-user mapping:

$$\begin{aligned} &\textit{LEFT} \rightarrow \textit{left} \\ &\textit{RIGHT} \rightarrow \textit{right} \\ &\textit{CENTER} \rightarrow \textit{news feed} \end{aligned} \tag{1}$$

This propagation can be done in different ways: (a) by training a proxy model for media and then applying it to users, (b) by additionally using label propagation (LP) for semi-supervised learning.

Let us describe the proxy model propagation for (a) first. Let M be the set of media, and U be the set of users. We say a user $u \in U$ mentions a medium $m \in M$ if u posts a tweet that contains a link to the website of m. We denote the set of users that mention the medium m as $C_m \subseteq U$.

We can therefore create a representation for a medium by aggregating the embeddings of the users that mention the target medium. Such a representation is convenient as it lies in the same space as the user representation. In particular, given a medium $m \in M$, we compute its representation $R(m)$ as

$$R(m) = \frac{1}{|C_m|} \sum_{u \in C_m} R(u), \tag{2}$$

where $R(u)$ is the representation of user u, i.e., one (or a concatenation) of the embeddings described in Section 3.1.

Finally, we can train a LR model that uses $R(m)$ as features and the label for the medium $l(m)$. This model can be applied to predict the label of a user u by using the same type of representation $R(u)$, and the label mapping in Equation 1.

Label Propagation (b) is a transductive, graph-based, semi-supervised machine learning algorithm that, given a small set of labeled examples, assigns labels to previously unlabeled examples. The labels of each example change in relationship to the labels of *neighboring* ones in a properly-defined graph.

More formally, given a partially-labeled dataset of examples $X = X_u \cup X_l$, of which X_l are labeled examples with labels Y_l, and X_u are unlabeled examples, and a similarity graph $G(X, E)$, the label propagation algorithm finds the set of unknown labels Y_u such that the number of discordant pairs $(u, v) \in E : y_u \neq y_v$ is minimized, where y_z is the label assigned to example z.

Role	Users	Tweets	User Example	Tweet Example
Left	233	427 141	@samirgooden	@MichaelSkolnik @KatrinaPierson @samesfandiari Trump folks need to stop going on CNN.
Right	630	711 668	@chirrmorre	BREAKING: Trump ERASES Obama's Islamic Refugee Policy! https://t.co/uPTneTMNM5
News Feed	54	598 226	@dailysandiego	Exit poll: Wisconsin GOP voters excited, scared about Trump #politics

Table 1: Statistics and examples from the IRA Russian Trolls Tweets dataset.

The algorithm works as follows: At every iteration of propagation, each unlabeled node updates its label to the most frequent one among its neighbors. LP reaches convergence when each node has the same label as the majority of its neighbors. We define two different versions of LP by creating two different versions of the similarity graph G.

LP1 *Label Propagation using direct mention.*
In the first case, the set of edges among users U in the similarity graph G consists of the logical OR between the 2-hop closure of the U2H and the U2M graph. That is, for each two users $u, v \in U$, there is an edge in the similarity graph $(u, v) \in E$ if u and v share a common hashtag or a common user mention

$$(u,h) \in \text{U2H} \wedge (v,h) \in \text{ U2H} \vee \\ (u,w) \in \text{ U2M } \wedge (v,w) \in \text{ U2M}$$

The graph therefore uses the same information that is available to the embeddings.

To this graph, which currently encompasses only the set of users U, we add connections to the set of media M. We add an edge between each pair (u, m) if $u \in C_m$. Then, we run the label propagation algorithm, which propagates the labels from the labeled nodes M to the unlabeled nodes U, thanks to the mapping from Equation 1.

LP2 *Label Propagation based on a similarity graph.*
In this case, we use the same representation for the media as in the proxy model case above, as described by Equation 2. Then, we build a similarity graph among media and users based on their embeddings. For each pair $x, y \in U \cup M$ there is an edge in the similarity graph $(x, y) \in E$ iff

$$\text{sim}(R(x), R(y)) > \tau,$$

where sim is a similarity function between vectors, e.g., cosine similarity, and τ is a user-specified parameter that regulates the sparseness of the similarity graph.

Finally, we perform label propagation on the similarity graph defined by the embedding similarity, with the set of nodes corresponding to M starting with labels, and with the set of nodes corresponding to U starting without labels.

4 Data

4.1 IRA Russian Troll Tweets

Our main dataset contains 2 973 371 tweets by 2848 Twitter users, which the US House Intelligence Committee has linked to the Russian Internet Research Agency (IRA). The data was collected and published by Linvill and Warren (2018), and then made available online.[1] The time span covers the period from February 2012 to May 2018.

The trolls belong to the following manually assigned roles: Left Troll, Right Troll, News Feed, Commercial, Fearmonger, Hashtag Gamer, Non English, Unknown. Kim et al. (2019) have argued that the first three categories are not only the most frequent, but also the most interesting ones. Moreover, focusing on these troll types allows us to establish a connection between troll types and the political bias of the news media they mention. Table 1 shows a summary of the troll role distribution, the total number of tweets per role, as well as examples of troll usernames and tweets.

4.2 Media Bias/Fact Check

We use data from Media Bias/Fact Check (MBFC)[2] to label news media sites. MBFC divides news media into the following bias categories: Extreme-Left, Left, Center-Left, Center, Center-Right, Right, and Extreme-Right. We reduce the granularity to three categories by grouping Extreme-Left and Left as LEFT, Extreme-Right and Right as RIGHT, and Center-Left, Center-Right, and Center as CENTER.

[1] http://github.com/fivethirtyeight/russian-troll-tweets

[2] http://mediabiasfactcheck.com

Bias	Count	Example
LEFT	341	www.cnn.com
RIGHT	619	www.foxnews.com
CENTER	372	www.apnews.com

Table 2: Summary statistics about the Media Bias/Fact Check (MBFC) dataset.

Table 2 shows some basic statistics about the resulting media dataset. Similarly to the IRA dataset, the distribution is right-heavy.

5 Experiments and Evaluation

5.1 Experimental Setup

For each user in the IRA dataset, we extracted all the links in their tweets, we expanded them recursively if they were shortened, we extracted the domain of the link, and we checked whether it could be found in the MBFC dataset. By grouping these relationships by media, we constructed the sets of users C_m that mention a given medium $m \in M$.

The U2H graph consists of $108\,410$ nodes and $443\,121$ edges, while the U2M graph has $591\,793$ nodes and $832\,844$ edges. We ran node2vec on each graph to extract 128-dimensional vectors for each node. We used these vectors as features for the fully supervised and for the distant-supervision scenarios. For Label Propagation, we used an empirical threshold for edge materialization $\tau = 0.55$, to obtain a reasonably sparse similarity graph.

We used two evaluation measures: accuracy, and macro-averaged F1 (the harmonic average of precision and recall). In the supervised scenario, we performed 5-fold cross-validation. In the distant-supervision scenario, we propagated labels from the media to the users. Therefore, in the latter case the user labels were only used for evaluation.

5.2 Evaluation Results

Table 3 shows the evaluation results. Each line of the table represents a different combination of features, models, or techniques. As mentioned in Section 3, the symbol '$\|$' denotes a single model trained on the concatenation of the features, while the symbol '$\oplus$' denotes an averaging of individual models trained on each feature separately. The tags 'LP1' and 'LP2' denote the two label propagation versions, by mention and by similarity, respectively.

We can see that accuracy and macro-averaged F1 are strongly correlated and yield very consistent rankings for the different models. Thus, henceforth we will focus our discussion on accuracy.

We can see in Table 3 that it is possible to predict the roles of the troll users by using distant supervision with relatively high accuracy. Indeed, the results for T2 are lower compared to their T1 counterparts by only 10 and 20 points absolute in terms of accuracy and F1, respectively. This is impressive considering that the models for T2 have no access to labels for troll users.

Looking at individual features, for both T1 and T2, the embeddings from U2M outperform those from U2H and from BERT. One possible reason is that the U2M graph is larger, and thus contains more information. It is also possible that the social circle of a troll user is more indicative than the hashtags they used. Finally, the textual content on Twitter is quite noisy, and thus the BERT embeddings perform slightly worse when used alone.

All our models with a single type of embedding easily outperform the model of Kim et al. (2019). The difference is even larger when combining the embeddings, be it by concatenating the embedding vectors or by training separate models and then combining the posteriors of their predictions.

By concatenating the U2M and the U2H embeddings (U2H $\|$ U2M), we fully leverage the hashtags and the mention representations in the latent space, thus achieving accuracy of 88.7 for T1 and 78.0 for T2, which is slightly better than when training separate models and then averaging their posteriors (U2H $\oplus$ U2M): 88.3 for T1 and 77.9 for T2. Adding BERT embeddings to the combination yields further improvements, and follows a similar trend, where feature concatenation works better, yielding 89.2 accuracy for T1 and 78.2 for T2 (compared to 89.0 accuracy for T1 and 78.0 for T2 for U2H $\oplus$ U2M $\oplus$ BERT).

Adding label propagation yields further improvements, both for LP1 and for LP2, with the latter being slightly superior: 89.6 vs. 89.3 accuracy for T1, and 78.5 vs. 78.3 for T2.

Overall, our methodology achieves sizable improvements over previous work, reaching an accuracy of 89.6 vs. 84.0 of Kim et al. (2019) in the fully supervised case. Moreover, it achieves 78.5 accuracy in the distant supervised case, which is only 11 points behind the result for T1, and is about 10 points above the majority class baseline.

Method	Full Supervision (T1)		Distant Supervision (T2)	
	Accuracy	Macro F1	Accuracy	Macro F1
Baseline (majority class)	68.7	27.1	68.7	27.1
Kim et al. (2019)	84.0	75.0	N/A	N/A
BERT	86.9	83.1	75.1	60.5
U2H	87.1	83.2	76.3	60.9
U2M	88.1	83.9	77.3	62.4
U2H $\oplus$ U2M	88.3	84.1	77.9	64.1
U2H $\|$ U2M	88.7	84.4	78.0	64.6
U2H $\oplus$ U2M $\oplus$ BERT	89.0	84.4	78.0	65.0
U2H $\|$ U2M $\|$ BERT	89.2	84.7	78.2	65.1
U2H $\|$ U2M $\|$ BERT + LP1	89.3	84.7	78.3	65.1
U2H $\|$ U2M $\|$ BERT + LP2	89.6	84.9	78.5	65.7

Table 3: Predicting the role of the troll users using full vs. distant supervision.

6 Discussion

6.1 Ablation Study

We performed different experiments with the hyper-parameters of the graph embeddings. With smaller dimensionality (i.e., using 16 dimensions instead of 128), we noticed 2–3 points of absolute decrease in accuracy across the board.

Moreover, we found that using all of the data for learning the embeddings was better than focusing only on users that we target in this study, namely *left*, *right*, and *news feed*, i.e., using the rest of the data adds additional context to the embedding space, and makes the target labels more contextually distinguishable. Similarly, we observe 5–6 points of absolute drop in accuracy when training our embeddings on tweets by trolls labeled as *left*, *right*, and *news feed*.

6.2 Comparison to Full Supervision

Next, we compared to the work of Kim et al. (2019), who had a fully supervised learning scenario, based on Tarde's Actor-Network Theory. They paid more attention to the content of the tweet by applying a text-distance metric in order to capture the semantic distance between two sequences. In contrast, we focus on critical elements of information that are salient in Twitter: *hashtags* and *user mentions*. By building a connection between users, hashtags, and user mentions, we effectively filtered out the noise and we focused only on the most sensitive type of context, thus automatically capturing features from this network via graph embeddings.

Method	Accuracy	Macro F1
Baseline (majority)	46.5	21.1
BERT	61.8	60.4
U2H	61.6	60.0
U2M	62.7	61.4
U2H $\oplus$ U2M	63.5	61.8
U2H $\|$ U2M	63.8	61.9
U2H $\oplus$ U2M $\oplus$ BERT	63.7	61.8
U2H $\|$ U2M $\|$ BERT	64.0	62.2

Table 4: Leveraging user embeddings to predict the bias of the media cited by troll users.

6.3 Reverse Classification: Media from Trolls

Table 4 shows an experiment in distant supervision for reverse classification, where we trained a model on the IRA dataset with the troll labels, and then we applied that model to the representation of the media in the MBFC dataset, where each medium is represented as the average of the embeddings of the users who cited that medium. We can see that we could improve over the baseline by 20 points absolute in terms of accuracy and by 41 in terms absolute in terms of macro-averaged F1.

We can see in Table 4 that the relative ordering in terms or performance for the different models is consistent with that for the experiments in the previous section. This suggests that the relationship between trolls and media goes both ways, and thus we can use labels for media as a way to label users, and we can also use labels for troll users as a way to label media.

7 Conclusion and Future Work

We have proposed a novel approach to analyze the behavior patterns of political trolls according to their political leaning (*left* vs. *news feed* vs. *right*) using features from social media, i.e., from Twitter. We experimented with two scenarios: (*i*) supervised learning, where labels for trolls are provided, and (*ii*) distant supervision, where such labels are not available, and we rely on more common labels for news outlets cited by the trolls. Technically, we leveraged the community structure and the text of the messages in the online social network of trolls represented as a graph, from which we extracted several types of representations, i.e., embeddings, for the trolls. Our experiments on the "IRA Russian Troll" dataset have shown improvements over the state-of-the-art in the supervised scenario, while providing a compelling case for the distant-supervision scenario, which has not been explored before.[3]

In future work, we plan to apply our methodology to other political events such as *Brexit* as well as to other election campaigns around the world, in connection to which large-scale troll campaigns have been revealed. We further plan experiments with other graph embedding methods, and with other social media. Finally, the relationship between media bias and troll's political role that we have highlighted in this paper is extremely interesting. We have shown how to use it to go from the media-space to the user-space and vice-versa, but so far we have just scratched the surface in terms of understanding of the process that generated these data and its possible applications.

Acknowledgments

This research is part of the Tanbih project,[4] which aims to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading. The project is developed in collaboration between the Qatar Computing Research Institute, HBKU and the MIT Computer Science and Artificial Intelligence Laboratory.

Gianmarco De Francisci Morales acknowledges support from Intesa Sanpaolo Innovation Center. The funder had no role in the study design, in the data collection and analysis, in the decision to publish, or in the preparation of the manuscript.

[3] Our data and code are available at http://github.com/amatanasov/conll_political_trolls

[4] http://tanbih.qcri.org/

References

Kayode Sakariyah Adewole, Nor Badrul Anuar, Amirrudin Kamsin, Kasturi Dewi Varathan, and Syed Abdul Razak. 2017. Malicious accounts: Dark of the social networks. *Journal of Network and Computer Applications*, 79:41–67.

Abdullah Almaatouq, Erez Shmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. 2016. If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security*, 15(5):475–491.

Amy Binns. 2012. DON'T FEED THE TROLLS! Managing troublemakers in magazines' online communities. *Journalism Practice*, 6(4):547–562.

Zhan Bu, Zhengyou Xia, and Jiandong Wang. 2013. A sock puppet detection algorithm on virtual spaces. *Know.-Based Syst.*, 37:366–377.

Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *Proceedings of the IEEE Conference on Privacy, Security, Risk, and Trust, and the IEEE Conference on Social Computing*, SocialCom/PASSAT '11, pages 1–8, Boston, MA, USA.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, Hyderabad, India.

Cheng Chen, Kui Wu, Venkatesh Srinivasan, and Xudong Zhang. 2013. Battling the internet water army: Detection of hidden paid posters. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 116–120, Niagara Falls, ON, Canada.

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 IEEE Conference on Privacy, Security, Risk and Trust and of the IEEE Conference on Social Computing*, PASSAT/SocialCom '12, pages 71–80, Amsterdam, Netherlands.

Kirsti K Cole. 2015. "It's like she's eager to be verbally abused": Twitter, trolls, and (en) gendering disciplinary rhetoric. *Feminist Media Studies*, 15(2):356–358.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2015.

Fame for sale: efficient detection of fake Twitter followers. *Decision Support Systems*, 80:56–71.

Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, and Yelena Mejova. 2017. Seminar users in the Arabic Twitter sphere. In *Proceedings of the 9th International Conference on Social Informatics*, SocInfo '17, pages 91–108, Oxford, UK.

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.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '2019, pages 4171–4186, Minneapolis, MN, USA.

Ibrahim Elbeltagi and Gomaa Agag. 2016. E-retailing ethics and its impact on customer satisfaction and repurchase intention: A cultural and commitment-trust theory perspective. *Internet Research*, 26(1):288–310.

Michael Fire, Dima Kagan, Aviad Elyashar, and Yuval Elovici. 2014. Friend or foe? Fake profile identification in online social networks. *Social Network Analysis and Mining*, 4(1):1–23.

Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. 2016. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1):42–53.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016. Quantifying controversy in social media. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 33–42, San Francisco, CA, USA.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. The effect of collective attention on controversial debates on social media. In *Proceedings of the 9th International ACM Web Science Conference*, WebSci '17, pages 43–52, Troy, NY, USA.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the International World Wide Web Conference*, WWW '18, pages 913–922, Lyon, France.

Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '17, pages 267–276, Varna, Bulgaria.

Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, San Francisco, CA, USA.

Hong-Youl Ha, Joby John, J. Denise John, and Yongkyun Chung. 2016. Temporal effects of information from social networks on online behavior: The role of cognitive and affective trust. *Internet Research*, 26(1):213–235.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2016. In search of credible news. In *Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '16, pages 172–180, Varna, Bulgaria.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, CIKM '15, pages 1835–1838, Melbourne, Australia.

Li-An Ho, Tsung-Hsien Kuo, and Binshan Lin. 2012. How social identification and trust influence organizational online knowledge sharing. *Internet Research*, 22(1):4–28.

Meng-Hsiang Hsu, Li-Wen Chuang, and Cheng-Se Hsu. 2014. Understanding online shopping intention: the roles of four types of trust and their antecedents. *Internet Research*, 24(3):332–352.

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.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, New Orleans, LA, USA.

Andreas M Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017a. We built a fake news & click-bait filter: What happened next will blow your mind! In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '17, pages 334–343, Varna, Bulgaria.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017b. Fully automated fact checking using external sources. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '17, pages 344–353, Varna, Bulgaria.

Dongwoo Kim, Timothy Graham, Zimin Wan, and Marian-Andrei Rizoiu. 2019. Tracking the digital traces of Russian trolls: Distinguishing the roles and strategy of trolls on Twitter. *CoRR*, abs/1901.05228.

Edward C. S. Ku. 2012. Beyond price, how does trust encourage online group's buying intention? *Internet Research*, 22(5):569–590.

Srijan Kumar, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 857–866, Perth, Australia.

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.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

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.

Xiaodong Li, Xinshuai Guo, Chuang Wang, and Shengliang Zhang. 2016. Do buyers express their true assessment? Antecedents and consequences of customer praise feedback behaviour on Taobao. *Internet Research*, 26(5):1112–1133.

Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2018. Attributed social network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 30(12):2257–2270.

Darren L Linvill and Patrick L Warren. 2018. Troll factories: The internet research agency and state-sponsored agenda building. *Resource Centre on Media Freedom in Europe*.

Clare Llewellyn, Laura Cram, Robin L. Hill, and Adrian Favero. 2018. For whom the bell trolls: Shifting troll behaviour in the Twitter Brexit debate. *JCMS: Journal of Common Market Studies*, 57(5):1148–1164.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Point process modelling of rumour dynamics in social media. 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 518–523, Beijing, China.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, IJCAI '16, pages 3818–3824, New York, NY, USA.

Michael McCord and M. Chuah. 2011. Spam detection on Twitter using traditional classifiers. In *Proceedings of the 8th International Conference on Autonomic and Trusted Computing*, ATC '11, pages 175–186, Banff, Canada.

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.

Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi Georgiev, and Ivan Koychev. 2018. The dark side of news community forums: Opinion manipulation trolls. *Internet Research*, 28(5):1292–1312.

Todor Mihaylov and Preslav Nakov. 2016. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 399–405, Berlin, Germany.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 860–869, Minneapolis, MN, USA.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadjov, and James Glass. 2018. Fact checking in community forums. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI '18, pages 879–886, New Orleans, LA, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, Lake Tahoe, NV, USA.

Tanushree Mitra, Graham P. Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 126–145, Portland, OR, USA.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 767–776, New Orleans, LA, USA.

Michael J Moore, Tadashi Nakano, Akihiro Enomoto, and Tatsuya Suda. 2012. Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, 28(3):861–867.

Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 441–450, Seattle, WA, USA.

Preslav Nakov, Tsvetomila Mihaylova, Lluís Màrquez, Yashkumar Shiroya, and Ivan Koychev. 2017. Do not trust the trolls: Predicting credibility in community question answering forums. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '17, pages 551–560, Varna, Bulgaria.

Shirui Pan, Jia Wu, Xingquan Zhu, Chengqi Zhang, and Yang Wang. 2016. Tri-party deep network representation. *Network*, 11(9):12.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17, pages 1003–1012, Perth, Australia.

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.

Carlos Ruiz, David Domingo, Josep Lluís Micó, Javier Díaz-Noci, Koldo Meso, and Pere Masip. 2011. Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. *The International Journal of Press/Politics*, 16(4):463–487.

Jari Salo and Heikki Karjaluoto. 2007. A conceptual model of trust in the online environment. *Online Information Review*, 31(5):604–621.

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.

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

Pnina Shachaf and Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3):357–370.

Thamar Solorio, Ragib Hasan, and Mainul Mizan. 2014. Sockpuppet detection in Wikipedia: A corpus of real-world deceptive writing for linking identities. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC '14, pages 1355–1358, Reykjavik, Iceland.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1067–1077, Florence, Italy.

Scott Thacker and Mark D Griffiths. 2012. An exploratory study of trolling in online video gaming. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL)*, 2(4):17–33.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Mengzhou Zhuang, Geng Cui, and Ling Peng. 2018. Manufactured opinions: The effect of manipulating online product reviews. *Journal of Business Research*, 87:24 – 35.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3):1–29.