TexTrolls: Identifying Russian Trolls on Twitter from a Textual Perspective

Abstract

The online new emerging suspicious users, that usually are called trolls, are one of the main sources of hate, fake, and deceptive online messages. Some agendas are utilizing these harmful users to spread incitement tweets, and as a consequence, the audience get deceived. The challenge in detecting such accounts is that they conceal their identities which make them disguised in social media, adding more difficulty to identify them using just their social network information. Therefore, in this paper, we propose a text-based approach to detect the online trolls such as those that were discovered during the US 2016 presidential elections. Our approach is mainly based on textual features which utilize thematic information, and profiling features to identify the accounts from their way of writing tweets. We deduced the thematic information in a unsupervised way and we show that coupling them with the textual features enhanced the performance of the proposed model. In addition, we find that the proposed profiling features perform the best comparing to the textual features.

Introduction

Recent years have seen a large increase in the amount of disinformation and fake news spread on social media. False information was used to spread fear and anger among people, which in turn, provoked crimes in some countries. The US in the recent years experienced many similar cases during the presidential elections, such as the one commonly known as “Pizzagate" . Later on, Twitter declared that they had detected a suspicious campaign originated in Russia by an organization named Internet Research Agency (IRA), and targeted the US to affect the results of the 2016 presidential elections. The desired goals behind these accounts are to spread fake and hateful news to further polarize the public opinion. Such attempts are not limited to Twitter, since Facebook announced in mid-2019 that they

detected a similar attempt originating from UAE, Egypt and Saudi Arabia and targeting other countries such as Qatar, Palestine, Lebanon and Jordan. This attempt used Facebook pages, groups, and user accounts with fake identities to spread fake news supporting their ideological agendas. The automatic detection of such attempts is very challenging, since the true identity of these suspicious accounts is hidden by imitating the profiles of real persons from the targeted audience; in addition, sometimes they publish their suspicious idea in a vague way through their tweets' messages.

A previous work BIBREF0 showed that such suspicious accounts are not bots in a strict sense and they argue that they could be considered as “software-assisted human workers". According to BIBREF1, the online suspicious accounts can be categorized into 3 main types: Robots, Cyborgs, and Human Spammers. We consider IRA accounts as another new emerging type called trolls, which is similar to Cyborgs except that the former focuses on targeting communities instead of individuals.

In this work, we identify online trolls in Twitter, namely IRA trolls, from a textual perspective. We study the effect of a set of text-based features and we propose a machine learning model to detect them. We aim to answer three research questions: RQ1. Does the thematic information improve the detection performance?, RQ2. Can we detect IRA trolls from only a textual perspective? and RQ3. How IRA campaign utilized the emotions to affect the public opinions?

The rest of the paper is structured as follows. In the following section, we present an overview on the literature work on IRA trolls. In Section SECREF3, we describe how the used dataset was compiled. Section SECREF4 describes our proposed features for our approach. The experiments, results, and analyses are presented in Section SECREF5. Finally, we draw some conclusions and discuss possible future work on IRA trolls.

Related Work on IRA Trolls

After the 2016 US elections, Twitter has detected a suspicious attempt by a large set of accounts to influence the results of the elections. Due to this event, an emerging research works about the Russian troll accounts started to appear BIBREF2, BIBREF3, BIBREF0, BIBREF4, BIBREF5.

The research works studied IRA trolls from several perspectives. The work in BIBREF4 studied the links' domains that were mentioned by IRA trolls and how much they overlap with other links used in tweets related to "Brexit". In addition, they compare "Left" and "Right" ideological trolls in terms of the number of re-tweets they received, number of followers, etc, and the online propaganda strategies they used. The authors in BIBREF2 analyzed IRA campaign in both Twitter and Facebook, and they focus on the evolution of IRA paid advertisements on Facebook before and after the US presidential elections from a thematic perspective.

The analysis work on IRA trolls not limited only to the tweets content, but it also considered the profile description, screen name, application client, geo-location, timezone, and number of links used per each media domain BIBREF3. There is a probability that Twitter has missed some IRA accounts that maybe were less active than the others. Based on this hypothesis, the work in BIBREF0 built a machine learning model based on profile, language distribution, and stop-words usage features to detect IRA trolls in a newly sampled data from Twitter. Other works tried to model IRA campaign not only by focusing on the trolls accounts, but also by examining who interacted with the trolls by sharing their contents BIBREF6. Similarly, the work BIBREF5 proposed a model that made use of the political ideologies of users, bot likelihood, and activity-related account metadata to predict users who spread the trolls' contents.

Data

To model the identification process of the Russian trolls, we considered a large dataset of both regular users (legitimate accounts) and IRA troll accounts. Following we describe the dataset. In Table TABREF6

we summarizes its statistics.

Data ::: Russian Trolls (IRA)

We used the IRA dataset that was released by Twitter after identifying the Russian trolls. The original dataset contains $3,841$ accounts, but we use a lower number of accounts and tweets after filtering them. We focus on accounts that use English as main language. In fact, our goal is to detect Russian accounts that mimic a regular US user. Then, we remove from these accounts non-English tweets, and maintain only tweets that were tweeted originally by them. Our final IRA accounts list contains 2,023 accounts.

Data ::: Regular Accounts

To contrast IRA behaviour, we sampled a large set of accounts to represent the ordinary behaviour of accounts from US. We collected a random sample of users that they post at least 5 tweets between 1st of August and 31 of December, 2016 (focusing on the US 2016 debates: first, second, third and vice president debates and the election day) by querying Twitter API hashtags related to the elections and its parties (e.g #trump, #clinton, #election, #debate, #vote, etc.). In addition, we selected the accounts that have location within US and use English as language of the Twitter interface. We focus on users during the presidential debates and elections dates because we suppose that the peak of trolls efforts concentrated during this period.

The final dataset is totally imbalanced (2% for IRA trolls and 98% for the regular users). This class imbalance situation represent a real scenario. From Table TABREF6, we can notice that the number of total tweets of the IRA trolls is similar to the one obtained from the regular users. This is due to the fact that IRA trolls were posting a lot of tweets before and during the elections in an attempt to try to make

their messages reach the largest possible audience.

Textual Representation

In order to identify IRA trolls, we use a rich set of textual features. With this set of features we aim to model the tweets of the accounts from several perspectives.

Textual Representation ::: Thematic Information

Previous works BIBREF7 have investigated IRA campaign efforts on Facebook, and they found that IRA pages have posted more than $\sim $80K posts focused on division issues in US. Later on, the work in BIBREF2 has analyzed Facebook advertised posts by IRA and they specified the main themes that these advertisements discussed. Given the results of the previous works, we applied a topic modeling technique on our dataset to extract its main themes. We aim to detect IRA trolls by identifying their suspicious ideological changes across a set of themes.

Given our dataset, we applied Latent Dirichlet Allocation (LDA) topic modeling algorithm BIBREF8 on the tweets after a prepossessing step where we maintained only nouns and proper nouns. In addition, we removed special characters (except HASH "#" sign for the hashtags) and lowercase the final tweet. To ensure the quality of the themes, we removed the hashtags we used in the collecting process where they may bias the modeling algorithm. We tested multiple number of themes and we chose seven of them. We manually observed the content of these themes to label them. The extracted themes are: Police shootings, Islam and War, Supporting Trump, Black People, Civil Rights, Attacking Hillary, and Crimes. In some themes, like Supporting Trump and Attacking Hillary, we found contradicted opinions, in favor and against the main themes, but we chose the final stance based on the most representative hashtags and words in each of them (see Figure FIGREF11). Also, the themes Police Shooting and Crimes are similar,

but we found that some words such as: police, officers, cops, shooting, gun, shot, etc. are the most discriminative between these two themes. In addition, we found that the Crimes theme focuses more on raping crimes against children and women. Our resulted themes are generally consistent with the ones obtained from the Facebook advertised posts in BIBREF2, and this emphasizes that IRA efforts organized in a similar manner in both social media platforms.

Based on our thematic information, we model the users textual features w.r.t. each of these themes. In other words, we model a set of textual features independently for each of the former themes to capture the emotional, stance, and others changes in the users tweets.

For the theme-based features, we use the following features that we believe that they change based on the themes:

Emotions: Since the results of the previous works BIBREF2, BIBREF7 showed that IRA efforts engineered to seed discord among individuals in US, we use emotions features to detect their emotional attempts to manipulate the public opinions (e.g. fear spreading behavior). For that, we use the NRC emotions lexicon BIBREF9 that contains $\sim $14K words labeled using the eight Plutchik's emotions.

Sentiment: We extract the sentiment of the tweets from NRC BIBREF9, positive and negative.

Bad & Sexual Cues: During the manual analysis of a sample from IRA tweets, we found that some users use bad slang word to mimic the language of a US citizen. Thus, we model the presence of such words using a list of bad and sexual words from BIBREF10.

Stance Cues: Stance detection has been studied in different contexts to detect the stance of a tweet reply with respect to a main tweet/thread BIBREF11. Using this feature, we aim to detect the stance of the

users regarding the different topics we extracted. To model the stance we use a set of stance lexicons employed in previous works BIBREF12, BIBREF13. Concretely, we focus on the following categories: belief, denial, doubt, fake, knowledge, negation, question, and report.

Bias Cues: We rely on a set of lexicons to capture the bias in text. We model the presence of the words in one of the following cues categories: assertives verbs BIBREF14, bias BIBREF15, factive verbs BIBREF16, implicative verbs BIBREF17, hedges BIBREF18, report verbs BIBREF15. A previous work has used these bias cues to identify bias in suspicious news posts in Twitter BIBREF19.

LIWC: We use a set of linguistic categories from the LIWC linguistic dictionary BIBREF20. The used categories are: pronoun, anx, cogmech, insight, cause, discrep, tentat, certain, inhib, incl.

Morality: Cues based on the morality foundation theory BIBREF21 where words labeled in one of a set of categories: care, harm, fairness, cheating, loyalty, betrayal, authority, subversion, sanctity, and degradation.

Given V_i as the concatenation of the previous features vectors of a tweet$_i$, we represent each user's tweets by considering the average and standard deviation of her tweets' $V_{1,2,..N}$ in each theme j independently and we concatenate them. Mathematically, a user x final feature vector is defined as follows:

where given the jth theme, N_j is the total number of tweets of the user, V_{ij} is the ith tweet feature vector, $\overline{V_j}$ is the mean of the tweets' feature vectors. With this representation we aim at capturing the "Flip-Flop" behavior of IRA trolls among the themes (see Section SECREF33).

Textual Representation ::: Profiling IRA Accounts

As Twitter declared, although the IRA campaign was originated in Russia, it has been found that IRA trolls concealed their identity by tweeting in English. Furthermore, for any possibility of unmasking their identity, the majority of IRA trolls changed their location to other countries and the language of the Twitter interface they use. Thus, we propose the following features to identify these users using only their tweets text:

Native Language Identification (NLI): This feature was inspired by earlier works on identifying native language of essays writers BIBREF22. We aim to detect IRA trolls by identifying their way of writing English tweets. As shown in BIBREF19, English tweets generated by non-English speakers have a different syntactic pattern . Thus, we use state-of-the-art NLI features to detect this unique pattern BIBREF23, BIBREF24, BIBREF25; the feature set consists of bag of stopwords, Part-of-speech tags (POS), and syntactic dependency relations (DEPREL). We extract the POS and the DEPREL information using spaCy, an off-the-shelf POS tagger. We clean the tweets from the special characters and maintained dots, commas, and first-letter capitalization of words. We use regular expressions to convert a sequence of dots to a single dot, and similarly for sequence of characters.

Stylistic: We extract a set of stylistic features following previous works in the authorship attribution domain BIBREF27, BIBREF28, BIBREF29, such as: the count of special characters, consecutive characters and letters, URLs, hashtags, users' mentions. In addition, we extract the uppercase ratio and the tweet length.

Similar to the feature representation of the theme-based features, we represent each user's tweets by considering the average and standard deviation of her tweets' $V_{1,2,..N}$, given V_i as the concatenation of the previous two features vectors of a tweet$_i$. A user x final feature vector is defined as follows:

where N is her total number of tweets, V_i is the ith tweet feature vector, $\overline{V}$ is the

mean of her tweets feature vectors.

Experiments and Analysis ::: Experimental Setup

We report precision, recall and F1 score. Given the substantial class imbalance in the dataset, we use the macro weighted version of the F1 metric. We tested several classifiers and Logistic Regression showed the best F1$_{macro}$ value. We kept the default parameters values. We report results for 5-folds cross-validation.

Experiments and Analysis ::: Baselines

In order to evaluate our feature set, we use Random Selection, Majority Class, and bag-of-words baselines. In the bag-of-words baseline, we aggregate all the tweets of a user into one document. A previous work BIBREF30 showed that IRA trolls were playing a hashtag game which is a popular word game played on Twitter, where users add a hashtag to their tweets and then answer an implied question BIBREF31. IRA trolls used this game in a similar way but focusing more on offending or attacking others; an example from IRA tweets: "#OffendEveryoneIn4Words undocumented immigrants are ILLEGALS". Thus, we use as a baseline Tweet2vec BIBREF32 which is a a character-based Bidirectional Gated Recurrent neural network reads tweets and predicts their hashtags. We aim to assess if the tweets hashtags can help identifying the IRA tweets. The model reads the tweets in a form of character one-hot encodings and uses them for training with their hashtags as labels. To train the model, we use our collected dataset which consists of $\sim $3.7M tweets. To represent the tweets in this baseline, we use the decoded embedding produced by the model and we feed them to the Logistic Regression classifier.

IRA dataset provided by Twitter contains less information about the accounts details, and they limited to: profile description, account creation date, number of followers and followees, location, and account

language. Therefore, as another baseline we use the number of followers and followees to assess their identification ability (we will mention them as Network Features in the rest of the paper).

Experiments and Analysis ::: Results

Table TABREF32 presents the classification results showing the performance of each feature set independently. Generally, we can see that the thematic information improves the performance of the proposed features clearly (RQ1), and with the largest amount in the Emotions features (see $-_{themes}$ and $+_{themes}$ columns). This result emphasizes the importance of the thematic information. Also, we see that the emotions performance increases with the largest amount considering F1$_{macro}$ value; this motivates us to analyze the emotions in IRA tweets (see the following section).

The result of the NLI feature in the table is interesting; we are able to detect IRA trolls from their writing style with a F1$_{macro}$ value of 0.91. Considering the results in Table TABREF32, we can notice that we are able to detect the IRA trolls effectively using only textual features (RQ2).

Finally, the baselines results show us that the Network features do not perform well. A previous work BIBREF3 showed that IRA trolls tend to follow a lot of users, and nudging other users to follow them (e.g. by writing "follow me" in their profile description) to fuse their identity (account information) with the regular users. Finally, similar to the Network features, the Tweet2vec baseline performs poorly. This indicates that, although IRA trolls used the hashtag game extensively in their tweets, the Tweet2vec baseline is not able to identify them.

Experiments and Analysis ::: Analysis

Given that the Emotions features boosted the F1$_{macro}$ with the highest value comparing to the other

theme-based features, in Figure FIGREF34 we analyze IRA trolls from emotional perspective to answer RQ3. The analysis shows that the themes that were used to attack immigrants (Black People and Islam and War) have the fear emotion in their top two emotions. While on the other hand, a theme like Supporting Trump has a less amount of fear emotion, and the joy emotion among the top emotions.

Why do the thematic information help? The Flip-Flop behavior. As an example, let's considering the fear and joy emotions in Figure FIGREF34. We can notice that all the themes that used to nudge the division issues have a decreasing dashed line, where others such as Supporting Trump theme has an extremely increasing dashed line. Therefore, we manually analyzed the tweets of some IRA accounts and we found this observation clear, as an example from user x:

Islam and War: (A) @RickMad: Questions are a joke, a Muslim asks how SHE will be protected from Islamaphobia! Gmaffb! How will WE be protected from terrori…

Supporting Trump: (B) @realDonaldTrump: That was really exciting. Made all of my points. MAKE AMERICA GREAT AGAIN!

Figure FIGREF35 shows the flipping behaviour for user x by extracting the mean value of the fear and joy emotions. The smaller difference between the fear and joy emotions in the Islam and War theme for this user is due to the ironic way of tweeting for the user (e.g. the beginning of tweet A: "Questions are a joke"). Even though, the fear emotion is still superior to the joy. We notice a similar pattern in some of the regular users, although much more evident among IRA trolls.

To understand more the NLI features performance, given their high performance comparing to the other features, we extract the top important tokens for each of the NLI feature subsets (see Figure FIGREF37). Some of the obtained results confirmed what was found previously. For instance, the authors in

BIBREF19 found that Russians write English tweets with more prepositions comparing to native speakers of other languages (e.g. as, about, because in (c) Stop-words and RP in (a) POS in Figure FIGREF37). Further research must be conducted to investigate in depth the rest of the results.

Linguistic Analysis. We measure statistically significant differences in the cues markers of Morality, LIWC, Bias and Subjectivity, Stance, and Bad and Sexual words across IRA trolls and regular users. These findings presented in Table TABREF38 allows for a deeper understanding of IRA trolls.

False Positive Cases. The proposed features showed to be effective in the classification process. We are interested in understanding the causes of misclassifying some of IRA trolls. Therefore, we manually investigated the false positive tweets and we found that there are three main reasons: 1) Some trolls were tweeting in a questioning way by asking about general issues; we examined their tweets but we did not find a clear ideological orientation or a suspicious behaviour in their tweets. 2) Some accounts were sharing traditional social media posts (e.g. "http://t.co/GGpZMvnEAj cat vs trashcan"); the majority of the false positive IRA trolls are categorized under this reason. In addition, these posts were given a false theme name; the tweet in the previous example assigned to Attacking Hillary theme. 3) Lack of content. Some of the misclassified trolls mention only external links without a clear textual content. This kind of trolls needs a second step to investigate the content of the external links. Thus, we tried to read the content of these links but we found that the majority of them referred to deleted tweets. Probably this kind of accounts was used to "raise the voice" of other trolls, as well as, we argue that the three kinds of IRA trolls were used for "likes boosting".

Conclusion

In this work, we present a textual approach to detect social media trolls, namely IRA accounts. Due to the anonymity characteristic that social media provide to users, these kinds of suspicious behavioural

accounts have started to appear. We built a new machine learning model based on theme-based and profiling features that in cross-validation evaluation achieved a F1$_{macro}$ value of 0.94. We applied a topic modeling algorithm to go behind the superficial textual information of the tweets. Our experiments showed that the extracted themes boosted the performance of the proposed model when coupled with other surface text features. In addition, we proposed NLI features to identify IRA trolls from their writing style, which showed to be very effective. Finally, for a better understanding we analyzed the IRA accounts from emotional and linguistic perspectives.

Through the manually checking of IRA accounts, we noticed that frequently irony was employed. As a future work, it would be interesting to identify these accounts by integrating an irony detection module.