

TexTrolls: Identifying Trolls on Twitter with Textual and Affective Features

Bilal Ghanem^a, Davide Buscaldi^b and Paolo Rosso^a

^aUniversitat Politècnica de València, Valencia, Spain

^bLIPN - Université Sorbonne Paris Nord, France

Abstract

The growing suspicious online 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 accounts to spread incitement tweets, and as a consequence, the online users get deceived. The challenge in detecting such accounts is that they conceal their real identities, adding more difficulty to identify them using just their social network information. In this paper we propose affective and lexical information -based models to detect the online trolls such as those that were discovered during the US 2016 presidential elections. Our approach is mainly based on features that take into account topic information and profiling features to identify the accounts from their way of writing tweets. We inferred the topic information in an unsupervised way and we show that coupling them with the affective and lexical features enhanced the performance of the proposed models. We find that the proposed profiling features perform the best comparing to the other features. Our approach shows superior results in comparison to several strong baselines.

Keywords

Profiling trolls, Twitter, topic modelings

1. Introduction


Recent years have seen a large increase in the amount of disinformation and fake news spread on social media. False information has been 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”¹. 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 elections². 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

OHARS'20: Workshop on Online Misinformation- and Harm-Aware Recommender Systems, September 25, 2020, Virtual Event

EMAIL: bigha@doctor.upv.es (B. Ghanem); buscaldi@lipn.univ-paris13.fr (D. Buscaldi); proso@dsic.upv.es (P. Rosso)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://www.rollingstone.com/politics/politics-news/anatomy-of-a-fake-news-scandal-125877/>

²https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html

and Jordan³. 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 accounts is hidden by imitating profiles of real persons from the targeted audience; in addition, sometimes these accounts publish their suspicious idea in a vague way through their tweets' messages.

A previous work [1] 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 Clark et al. [2], the online suspicious accounts can be categorized into 3 main types: Robots, Cyborgs, and Human Spammers. We consider IRA accounts as the new emerging type called trolls, which is similar to Cyborgs except that the former focuses on targeting communities instead of individuals⁴.

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, including affective ones, and we propose machine learning models that take into account topic information. These models can be applied to go beyond the textual superficial features, that are used in the related works, to detect advanced online manipulating efforts. We also conduct an in-depth analysis of the trolls' language to provide evidence for the reader about their online manipulation campaigns. In this research work, we aim to answer three research questions:

RQ1 *Can we detect IRA trolls from only a textual perspective?*

RQ2 *Does the topic information improve the detection performance?*

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 3, we describe how the used dataset was compiled. Section 4 describes our proposed features for our approaches. The experiments and results are presented in Section 6. We present an analysis for the trolls campaign in Section 7. Finally, we discuss the limitations of the proposed models and we draw some conclusions and possible future work on the identification of trolls.

2. Related Work

2.1. IRA Trolls

After the 2016 US elections, Twitter detected a suspicious attempt by a large set of accounts to influence the results of the elections. Due to this event, various research works about the Russian troll accounts started to appear [3, 4, 1, 5, 6].

These research works studied IRA trolls from several perspectives, but most of them focused on analyzing them and studying their strategies, instead of building a detection model. The work in [5] 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,

³<https://newsroom.fb.com/news/2019/08/cib-uae-egypt-saudi-arabia/>

⁴<https://itstillworks.com/difference-between-troll-cyberbully-5054.html>

etc, and the online propaganda strategies they used. The authors in [3] 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 topic perspective, e.g. whose topics IRA trolls targeted to seed discord among the public.

The analysis works on IRA trolls were not limited only to analyse the tweets content, but they also considered profile description, screen name, application client, geo-location, timezone, and number of links used per each media domain [4]. 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 [1] (*Still Out There*) 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 [7]. Similarly, the work [6] 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.

2.2. Online Bots

Online social bots have been a source of nuisance for the social media users for their suspicious behaviour in retweeting duplicated tweets or boosting advertisements tweets. The work in [8] studied a large portion of Twitter bots collected during a study of seven months. The authors study the behaviour of these bots and they grouped them into a set of categories, e.g. duplicate spammers, malicious promoters, friend infiltrators.

Bots detection has gained the attention of the research community. Recently, a shared task on bots profiling in Twitter [9] has been organized at PAN-2019 Lab targeting both Spanish and English languages. The best performing system [10] for the English language obtained an accuracy value of ~96%. The system is based on stylistic features such as terms occurrence, tweets length, number of capitalized words, etc., and employed a Random Forest classifier. Another work [11] proposed a system called *SentiBot* for detecting Indian bots in Twitter. The approach uses a large combination of features but mainly focuses on sentiment features. The used features in the previous works were not limited to stylistic and sentiments features: the authors of [12] proposed a SOTA system called *Botometer*⁵ that uses sentiment, friend, content, user, temporal, and network features.

3. Data

To model the detection of the IRA trolls, we considered a large dataset of both regular users (legitimate accounts) and IRA troll accounts. In the following we describe the dataset. In Table 1 we summarize its statistics.

⁵<https://github.com/IUNetSci/botometer-python>

Table 1
Statistics of the dataset.

	IRA Trolls	Regular Accounts
Total # of Accounts	2,023	94,643
Total # of Tweets	~ 1.8 M	~ 1.9 M
Avg. # of Tweets	357	19
Avg. # of Followers	1,834	9,867
Avg. # of Followees	1,025	2,277

3.1. The Internet Research Agency Dataset

We used the IRA dataset⁶ 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 the 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.

3.2. 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 are located 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 1, 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.

4. Textual Features

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.

4.1. Topic Information

Previous works [13] have investigated IRA campaign efforts on Facebook, and they found that IRA pages have posted more than ~80K posts focused on divisive issues in US. Later on, the

⁶https://about.twitter.com/en_us/values/elections-integrity.html

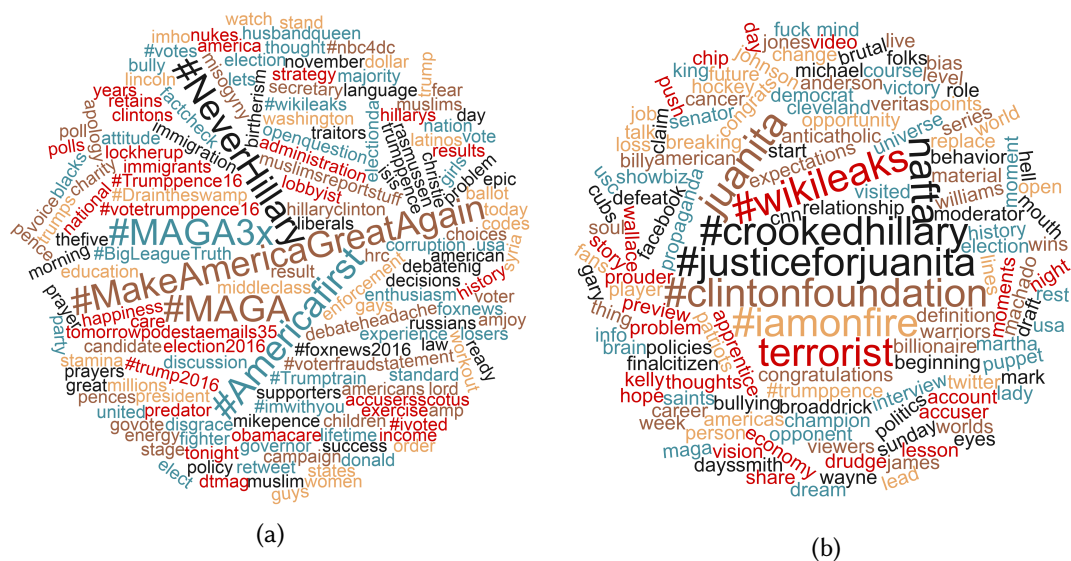


Figure 1: (a) *Trump* and (b) *Hillary* topics words clouds.

work in [3] has analyzed Facebook advertised posts by IRA and they specified the main topics that these advertisements discussed. Given the results of the previous works, we applied a topic modeling technique, namely Latent Dirichlet Allocation (LDA) [14], on our dataset to extract its main topics. We aim to detect IRA trolls by identifying their suspicious ideological changes across a set of topics.

Given our dataset, we applied LDA on the tweets after a preprocessing step where we maintained only nouns and proper nouns using the SpaCy part-of-speech (POS) tagger, which is an off-the-shelf POS tagger⁷. In addition, we removed special characters (except HASH “#” sign for the hashtags) and lowercase the final tweet. To ensure the quality of the topics, we removed the hashtags we used in the collecting process where they may bias the modeling algorithm. We tested multiple numbers of topics and finally we use seven. We manually observed the content of these topics to label them. The extracted topics (*T*) are: *Police shootings, Islam and War, Trump, Black People, Civil Rights, Hillary, and Crimes*. In some topics, like *Trump* and *Hillary*, we found contradicted opinions, in favor and against the main topics, but generally we can notice that the *Trump* topic has a support stance to Trump, on the other hand, the *Hillary* topic has an against stance towards Hillary (see Figure 1 for the frequency-based wordcloud). Also, the topics *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 topics. In addition, we found that the *Crimes* topic focuses more on raping crimes against children and women. Our resulted topics are generally consistent with the ones obtained from the Facebook advertised posts in [3], and this emphasizes that IRA efforts organized in a similar manner in both social media platforms.

Based on our topic information, we model the users textual features w.r.t. each of these topics.

⁷<https://spacy.io/models>

In other words, we model a set of textual features which could change in the users' tweets across the topics. We aim at modeling the trolls manipulating effort in which they interact in a different way with each topic; e.g. a troll account may trigger positive emotions in a set of topics in favor of and negative if against. Similarly, showing supporting stance intensively in some topics and denial stance in others. Thus, we used LDA to annotate the tweets of the users in one of the T topics to capture the changes of the following proposed features among the topics.

We chose the following affective and lexical features under the assumption that they may characterise the trolls' language changes across the topics. We use term frequency representation to extract the following features from the tweets:

- **Emotions:** Since the results of the previous works [3, 13] 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 behaviour). For that, we use the NRC emotions lexicon [15] that contains $\sim 14K$ words labeled using the eight Plutchik's emotions (*8 Features*).
- **Sentiment:** We extract the sentiment of the tweets from NRC [15], *positive* and *negative* (*2 Features*).
- **Bad & Sexual Cues:** During the manual analysis of a sample from IRA tweets, we found that some users use bad words 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 [16] (*2 Features*).
- **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 [17]. 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 [18, 19]. Concretely, we focus on the following categories: *belief, denial, doubt, fake, knowledge, negation, question, and report* (*8 Features*).
- **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* [20], *bias* [21], *factive verbs* [22], *implicative verbs* [23], *hedges* citehyland2018metadiscourse, *report verbs*. A previous work has used these bias cues to identify bias in suspicious news posts in Twitter [24] (*6 Features*).
- **LIWC:** We use a set of linguistic categories from the LIWC linguistic dictionary [25]. The used categories are: *pronoun, anx, cogmech, insight, cause, discrep, tentat, certain, inhib, incl*⁸ (*10 Features*).
- **Morality:** Cues based on the morality foundation theory [26] where words labeled in one of a set of categories: *care, harm, fairness, unfairness, loyalty, betrayal, authority, subversion, sanctity, and degradation* (*10 Features*).

⁸Total pronouns, Anxiety, Cognitive processes, Insight, Causation, Discrepancy, Tentative, Certainty, Inhibition, and Inclusive respectively.

4.2. 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, as well as, 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 [27]. We aim to detect IRA trolls by identifying their way of writing English tweets. As shown in [24], English tweets generated by non-English speakers have a different syntactic pattern. Thus, we use SOTA NLI features to detect this unique pattern [28, 29, 30]. The feature set consists of bag of: stopwords (*179 Features*), POS tags (*46 Features*), and syntactic dependency relations (DEPREL) (*45 Features*). We extract the POS and the DEPREL information using spaCy. To normalize the tweets, we clean them from the special characters and maintain 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 (in total *270 Features*).
- **Stylistic:** We extract a set of stylistic features following previous works in the authorship attribution domain [31, 32, 33], such as: the count of special characters, consecutive characters and letters⁹, URLs, hashtags, users’ mentions. In addition, we extract the uppercase ratio and the tweet length (*8 Features*).

5. Models

Given the two sets of features that we presented in Section 4, we use them in two different approaches in order to build trolls detectors. The proposed approaches utilize a classical machine learning classifier and a Convolutional Neural Network (CNN):

All Features + LG. In this approach, we model the extracted textual features as follows: Given V_n as the concatenation of the previous 46 topic information features of a tweet n , we represent each user by considering the *average* and *standard deviation* of her tweets’ $V_{1,2,..,N}$ in each topic t independently. We concatenate the final vectors; final vectors are seven since the number of topics (T) is equals seven in our case. Mathematically, the final feature vector of a user x is defined as follows:

$$user_x = \bigodot_{t=1}^T \left[\frac{\sum_{n=1}^{N_t} V_{nt}}{N_t} \odot \sqrt{\frac{\sum_{n=1}^{N_t} (V_{nt} - \bar{V}_t)^2}{N_t}} \right] \quad (1)$$

where given the t^{th} topic, N_t is the total number of tweets of the user (annotated with the t^{th} topic), V_{nt} is the n^{th} tweet feature vector, \bar{V}_t is the mean of the tweets’ feature vectors; \odot represents the vectors concatenation process. With this representation we aim at capturing the “Flip-Flop” behaviour of the IRA trolls among the topics (see Section 7).

⁹We considered 2 or more consecutive characters, and 3 or more consecutive letters.

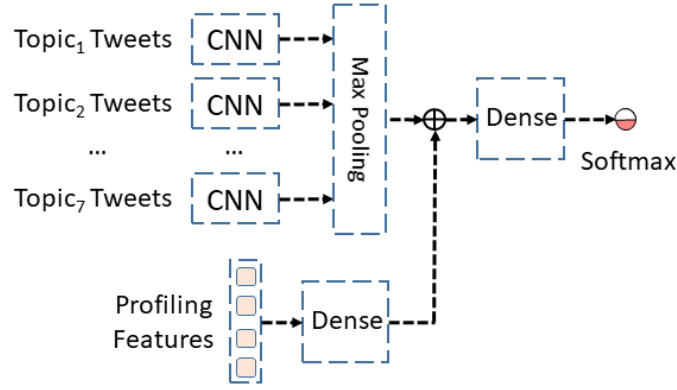


Figure 2: CNN structure.

Regarding the profiling features, we represent each user by considering the *average* and the *standard deviation* of her tweets' feature vectors, similar to the representation of the previous features but without considering the topic information. In short, we apply the *average* and the *standard deviation* on all the tweets of a user at once:

$$user_x = \frac{\sum_{n=1}^N V_n}{N} \odot \sqrt{\frac{\sum_{n=1}^N (V_n - \bar{V})^2}{N}} \quad (2)$$

where N is her total number of tweets, V_n is the n^{th} tweet feature vector, \bar{V} is the mean of the tweets feature vectors of a user x .

After preparing the two feature set vectors, we concatenate them, and we feed them to a Logistic Regression (LG) classifier.

CNN. We use a CNN to model the proposed features. We use a CNN that has two branches: one models the topic information (A) and the other models the profiling features (B). Figure 2 shows the proposed network.

In branch A, first we divide a user's tweets into seven tweets' groups based on their topics and then we feed each group to a different CNN. The tweets of a specific group are considered as one long document. Each CNN applies a convolution and max-pooling layers. The input document D of length n is represented as $[D_1, D_2, \dots, D_n]$ where $D_n \in \mathbb{R}^d$; \mathbb{R}^d is a d -dimensional one-hot vector of the i -th word in the input document. The words' d -dimensional vectors have a length of 46, that is, the total number of topic information features. After processing the input group of tweets, we apply another max-pooling layer to extract the important global features from the seven topics' CNNs. The structure of this branch is inspired by the Hierarchical Attention model [34] that has been proposed for document classification.

On the other hand, for branch B we concatenate all tweets of a user into one document, and we use the Equation 2 to extract a vector of the profiling features (length of 278) and we feed it to a dense layer $f(W_a v + b_a)$, where W_a and b_a are the corresponding weight matrix and bias terms, and f is an activation function, such as *ReLU*, *tanh*, etc.

After processing the input tweets in both branches, we concatenate the output vectors (\oplus)

and we feed them to another dense layer to learn their joint interaction. Finally, to get the classes probability of a document, we add a Softmax layer.

6. Experiments and Results

6.1. Experimental Setup

Given the substantial class imbalance in the dataset, we report precision, recall and F1 metrics on the IRA trolls (positive class). In the following section, we tested several classifiers¹⁰ with some of our baselines, and we highlight the ones that obtain the best F1 value. We kept the default parameters values. We report results for 5-folds cross-validation. Regarding the CNN model, after we divide the tweets into 7 groups in the branch A, we set their maximum length to 500 words and we pad the shorter ones with zeros. We noticed that some users have no tweets labeled with some topics. Thus, we substitute missing topic groups with zeros vectors. For hyper-parameter selection, we slice a 0.2 of the data for the validation part and we apply cross-validation on the rest. We tune various parameters with the corresponding search spaces: the sizes of the dense layers (32, 64, 128, 256), activation functions (*tanh*, *ReLU*), CNN filters' sizes (*different combinations of the sizes 3,4,5,6*) and their numbers (32, 64, 128, 256, 512), and the optimization function (*Adam*, *RMSprop*, *SGD*). Also, it is worthy of mentioning that we tried to oversample the minority class to improve the performance by randomly replicating the trolls users. However, we did not notice a clear improvement in the F1 metric.

6.2. Baselines

In order to evaluate our approach, we use the following baselines:

BOW + LR: We use bag-of-words (BOW) representation (weighted using TF-IDF scheme) with a LR classifier where we aggregate all the tweets of a user into one long document. We aim to assess how a simple word-based model can perform.

LSTM: Word embeddings-based models showed significant improvements in many tasks previously. We use Long short-term memory (LSTM) network [35] with *Glove (840b.300d) words embeddings* [36]. Similar to BOW baseline, we we aggregate all the tweets of a user into one long document.

Number of Tweets + NB: Based on the dataset statistics (see Table 1), we can notice that the IRA accounts have a large amount of tweets. Thus, as a baseline, we use the number of tweets for each account and we feed them to a NB classifier. We use this baseline to investigate if it is possible to detect the trolls accounts using only the number of tweets.

Tweet2vec + LR: A previous work [37] 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 [38]. IRA trolls used this game in a similar way but focusing more on offending or attacking the targeted section of the audience; an example from IRA tweets:

¹⁰We tested Logistic Regression (LR), Random Forest (RF) with 100 as the number of estimators, Naive Bayes (NB), Support Vector Machine (SVM) with its both kernels, and Neural Network (NN) with a single hidden layer of size 50 and *tanh* as an activation function.

Thus, we use as a baseline *Tweet2vec* [39] which is a character-based Bidirectional Gated Recurrent neural network that 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 ~ 3.7 M tweets¹¹. To represent the tweets in this baseline, we use the decoded embedding produced by the model and we feed them to a LR classifier.

Network Features + LR: IRA dataset provided by Twitter contains few information about the accounts details, and they are limited to: profile description, account creation date, number of followers and followees, location, and account language. Therefore, as a baseline we use the number of followers and followees to assess their identification performance. We feed these features to a LR classifier.

Botometer + RF: *Botometer* is the SOTA bots detection system, which uses content, sentiment, friend, network, temporal, and user features. We extract these features and feed them to a Random Forest (RF) classifier with 100 as the number of estimators following the authors setup.

Still Out There + ABDT: Also as a baseline, we use the available proposed model in the related work [1], which uses profile, language distribution, and stop-words usage features with an Adaptive Boosted Decision Trees (ABDT) classifier.

6.3. Results

Table 2 presents the classification results of the baselines and our approaches. We report the results of our classical classifier -based approach with top 3 performing classifiers (RF, NN, and LR). The best results in terms of F1 score obtained with the LR classifier. The results show that both proposed models perform best comparing to the used baselines. Also, the results show that the *All Features + LG* model performs better than the *CNN* with a noticeable difference in terms of F1 measure. Generally, we can notice that we are able to detect the IRA trolls effectively using the the affective and lexical features (RQ1).

The *topic* features have a good performance comparing to most of the baselines. The result obtained with the *Profiling* features is interesting; we are able to detect the IRA trolls from the users' writing style with an F1 value of 0.88 using the *All Features + LG* model. To assess whether the topic information improves the performance of each of the lexical features, we run the *All Features + LG* model with each feature independently, with and without utilizing the topic information (without considering the topics in Eq. 1). Following, we present the results obtained with each feature: Emotions (+0.74|-0.02)¹²; Sentiment (+0.28|-0.0); Bad & Sexual (+0.58|-0.0); Stance Cues (+0.72|-0.12); Bias Cues (+0.73|-0.03); LIWC (+0.71|-0.04), and Morality (+0.72|-0.36). We conclude from these results that the model weakly detects the changes in stances, variations in emotions, etc., for a user when we discard the topic information. Clearly, we can notice that the model became aware of the flipping behaviour across the topics. These results emphasize the importance of the topic information (RQ2), especially with the emotions.

¹¹We used the default parameters that were provided with the system code.

¹²(+) stands for the F1 result with the topic information and (-) without them.

Table 2
Classification results.

Method	Precision	Recall	F1
Network Features + LR	0.0	0.0	0.0
Random Selection	0.02	0.5	0.04
<i>Tweet2vec</i> + LR	0.18	0.64	0.28
Number of Tweets + NB	0.47	0.53	0.5
BOW + LR	0.86	0.51	0.64
LSTM	0.86	0.69	0.76
<i>Still Out There</i> + ABDT [1]	0.97	0.75	0.84
<i>Botometer</i> + RF	0.99	0.76	0.86
Topic Information Features			
Topic-based Features + LR	0.89	0.7	0.78
CNN (branch A)	0.79	0.81	0.80
Profiling Features			
Profiling Features + LR	0.92	0.85	0.88
CNN (branch B)	0.81	0.88	0.84
All Features			
All Features + RF	0.99	0.78	0.88
All Features + NN	0.90	0.89	0.90
All Features + LR	0.93	0.88	0.91
CNN	0.86	0.90	0.88

This motivates us to analyze further the emotions in the IRA tweets (see the following section). Finally, the baselines’ results show us that the *Network features* are not able to detect the IRA trolls. A previous work [4] showed that the IRA trolls tend to follow many users, and nudging other users to follow them (e.g. by writing “follow me” in their profile description) to hide their identity (account information) with the regular users. Finally, similar to the *Network features*, the *Tweet2vec* baseline performs poorly. This indicates that, although the IRA trolls used the hashtag game extensively in their tweets, the *Tweet2vec* baseline is not able to identify them. The results of both *Botometer* and *Still Out There* [1] are superior to the other baselines, but still lower comparing to our proposed approaches.

7. Analysis

Given that the *Emotions* feature boosted the F1 with the highest value comparing to the other topic-based features, in Figure 3 we analyze IRA trolls from an emotional perspective to answer RQ3. This analysis can make us more aware of the manipulation efforts across the topics. The figure shows that the topics that were used to attack immigrants (*Black People* and *Islam and War*) have the *fear* emotion in their top two emotions. On the other hand, a topic like *Trump* has the lowest amount of *fear* emotion, while the *joy* emotion is among the top emotions. *Why do the topic information help?..The Flip-Flop behaviour.* The trolls accounts were supporting their ideologies by tweeting positively in some topics, and simultaneously, posting tweets with a negative stance in topics that are against their ideologies. As an example, let’s considering the *fear* and *joy* emotions in Figure 3. We can notice that all the topics that used to nudge

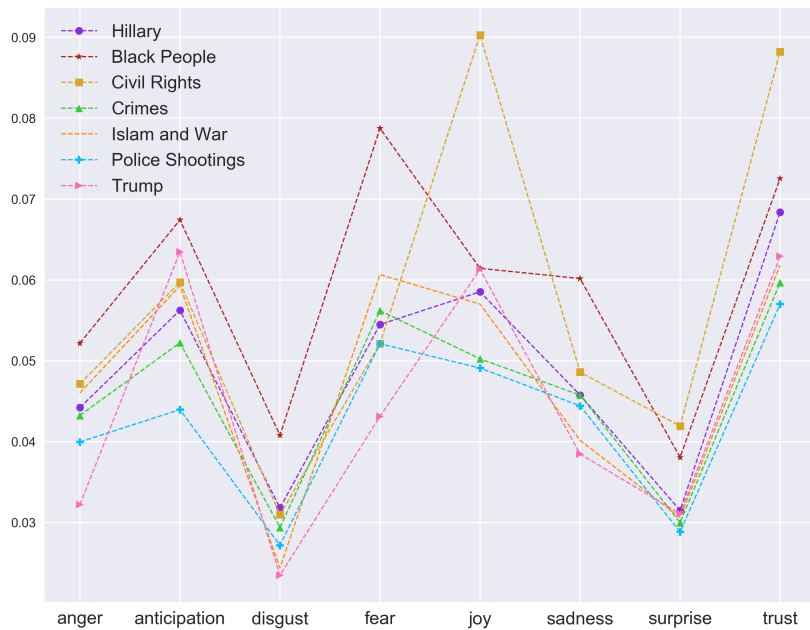


Figure 3: Emotional analyzing of the IRA trolls from an thematic perspective.

the divisive issues have a decreasing dashed line, where others such as *Trump* topic has an extremely increasing dashed line. Therefore, we manually analyzed the tweets of a sample of the IRA accounts and we found this observation clear, as an example from user *x*:

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

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

Figure 4 shows the flipping behaviour for user *x* by extracting the mean value of the *fear* and *joy* emotions. The small difference between *fear* and *joy* emotions in the *Islam and War* topic 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 noticed a similar pattern with some of the regular users, although much more evident among the IRA trolls. Thus, the way we combine our feature set with the topic information makes our classification models aware of this flipping behaviour.

To understand more the *NLI features* performance, given their high performance comparing to the other feature set, we extract the top important tokens for each of the *NLI* feature subsets (see Figure 5). Some of the obtained results confirmed what was found previously. For instance, the authors in [24] 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 RP¹³ in (a)

¹³RP stands for “adverb, particle” in the POS tag set.

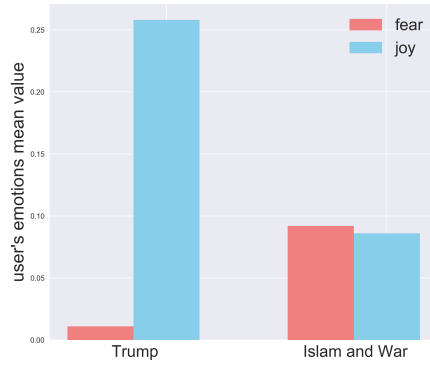


Figure 4: Flipping emotions between topics by user x (an IRA troll account).

Table 3

Linguistic analysis of Morality, LIWC, Bias and Subjectivity, Stance, and Bad and Sexual cues shown as the **percentage of averaged value of tweets with one or more cues** across IRA trolls (X) and regular users (Y) in a shape of X(arrows)Y. The tweets average value is the mean value across the topics. We report only significant differences: $p\text{-value} \leq 0.001 \uparrow \uparrow \uparrow$, $\leq 0.01 \uparrow \uparrow$, $\leq 0.05 \uparrow$ estimated using the Mann-Whitney U test. NSD stands for No statistical Significant Difference.

Morality		LIWC		Bias language		Stance		Bad and Sexual	
category	P_{value}	category	P_{value}	category	P_{value}	category	P_{value}	category	P_{value}
care	1.3 $\uparrow \uparrow \uparrow$.74	pronoun	53.34 $\downarrow \downarrow \downarrow$ 47.59	assertive	6.53 $\downarrow \downarrow \downarrow$ 7.05	belief	2.9 $\uparrow \uparrow \uparrow$.49	bad	5.4 $\uparrow \uparrow \uparrow$.66
harm	2.3 $\uparrow \uparrow \uparrow$.61	anx	1.9 $\uparrow \uparrow \uparrow$.98	bias	NSD	denial	0.6 $\uparrow \uparrow \uparrow$.57	sexual	3.5 $\uparrow \uparrow \uparrow$.16
fairness	0.64 $\downarrow \downarrow \downarrow$ 0.84	cogmech	NSD	factive	5.5 $\uparrow \uparrow \uparrow$.95	doubt	1.3 $\uparrow \uparrow \uparrow$.25	-	-
unfairness	0.06 $\downarrow \downarrow \downarrow$ 0.31	insight	12.1 $\uparrow \uparrow \uparrow$ 0.08	hedge	10.0 $\uparrow \uparrow \uparrow$.69	fake	0.49 $\downarrow \downarrow \downarrow$ 1.22	-	-
loyalty	0.84 $\downarrow \downarrow \downarrow$ 1.26	cause	10.7 $\uparrow \uparrow \uparrow$ 0.27	implicative	9.0 $\uparrow \uparrow \uparrow$.37	knowledge	0.75 $\downarrow \downarrow \downarrow$ 1.48	-	-
betrayal	0.13 $\downarrow \downarrow \downarrow$ 0.35	discrep	12.7 $\uparrow \uparrow \uparrow$ 1.07	report	14.37 $\downarrow \downarrow \downarrow$ 18.89	negation	11.4 $\uparrow \uparrow \uparrow$.10	-	-
authority	1.59 $\downarrow \downarrow \downarrow$ 1.88	tentat	13.9 $\uparrow \uparrow \uparrow$ 2.29	strong subj	54.1 $\uparrow \uparrow \uparrow$ 9.9	question	3.1 $\uparrow \uparrow \uparrow$.44	-	-
subversion	0.3 $\downarrow \downarrow \downarrow$ 1.33	certain	13.5 $\uparrow \uparrow \uparrow$ 0.69	weak subj	50.33 $\uparrow \uparrow \uparrow$ 41.96	report	2.86 $\downarrow \downarrow \downarrow$ 3.46	-	-
sanctity	0.4 $\uparrow \uparrow \uparrow$.27	inhib	4.1 $\uparrow \uparrow \uparrow$.87	-	-	-	-	-	-
degradation	0.5 $\uparrow \uparrow \uparrow$.49	incl	20.69 $\downarrow \downarrow \downarrow$ 21.24	-	-	-	-	-	-

POS in Figure 5). 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 3 allow for a deeper understanding of IRA trolls language usage. In general, the table shows that most of the topic-based features have a significant difference between the trolls and the regular users. Also, the analysis shows that trolls have a higher percentage of usage of Subjective language, Discrepancy, Bad, and Sexual terms comparing to the regular users. On the other hand, trolls appear to be less Fair (Fairness) and Loyal, and in addition use less Assertive and Report terms. Other categories like Anxiety and Bias do not show any significant difference. Considering that the Bias category has no significant difference emphasizes the advancement of the IRA campaign which was able to conceal its bias in text. Our approaches uses the topic information to overcome the limitations of the text only.

Topic Importance. The topics that were targeted by the Russian trolls are not equally important. They received tweets from both troll and regular users, but some of them have received

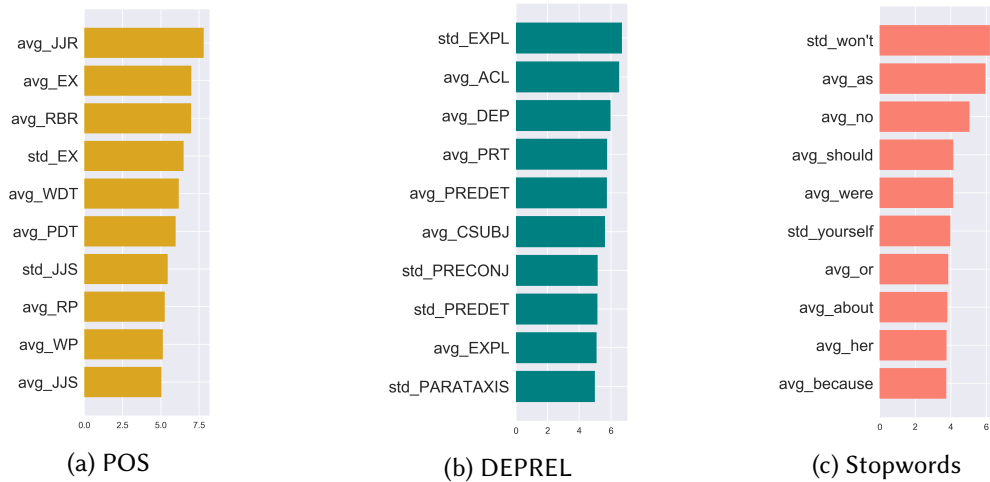


Figure 5: The top 10 important tokens in each of the NLI features.

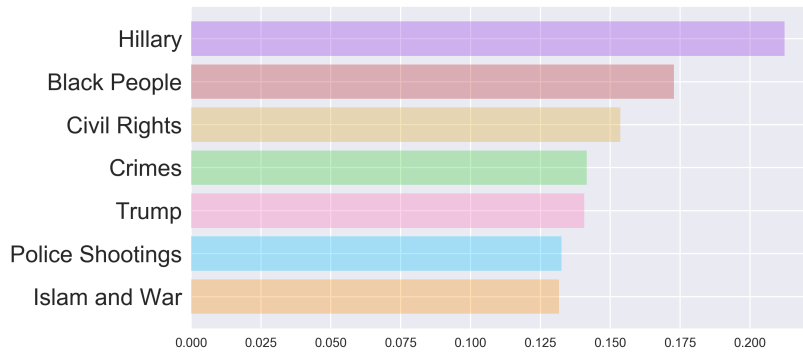


Figure 6: Topics importance.

more tweets from the trolls than the regular users. In this experiment, we extract the topics' importance to understand clearly the trolls campaign. For that, we extract the features importance values from our classifier, and then we average these values for each topic independently, given that each topic has the same feature set (see Eq. 1). Using these averaged values, we are able to rank the topics from the most important to the less important one in the classification process (see Figure 6). On one hand, we can notice that topics like *Attacking Hillary* and *Black People* are more important comparing to the others. The trolls targeted them more to reach their desired goal. On the other hand, topics like *Islam and War* and *Police Shootings* have the lowest importance; this could be justified by that regular users are most likely to have a high effect on these topics compared to the rest. Thus, these two topics became less important for the classifier to discriminate between the two types of accounts.

For understanding the language usages in each topic, we extract the *top 3* important features wrt each topic in Figure 7. The results obtained are consistent with our preliminary hypothesis about the trolls' language in each topic. The results show, for example, that the topic *Police*

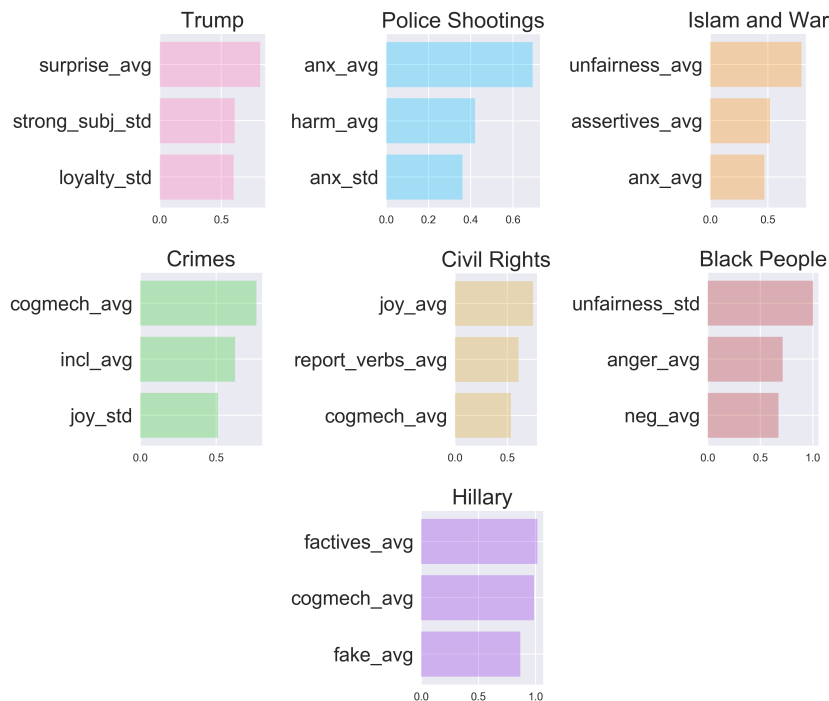


Figure 7: Top 3 features in the extracted topics.

Shooting has tweets that trigger at most *Anxiety* and *Harm*. The aim of these tweets is to increase the amount of panic among the public. Similarly, the *Black People* topic shows that it has *Unfairness*, *Anger*, and *Negative* emotions. This result is interesting and reveals the negative stance in the racist tweets against black people.

False Negative Cases. The proposed features showed to be effective in the classification processes. We are interested in understanding the causes of misclassifying some of IRA trolls. Therefore, we manually investigated the tweets of some of the false negative users 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 topic name; the tweet in the previous example assigned to *Attacking Hillary* topic. 3) Lack of content. Some of the misclassified trolls mentioned 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”.

8. Limitations

In this work, we focused on the detection of online trolls, namely the IRA Russian trolls. We proposed a topic and profiling -based approaches, and we compared them to several solid baselines. For the topic-based features, we used a set of lexicons as features and we combined them with topic information. This combination allowed our approaches to detect the change in the affective and lexical information among the extracted topics, and consequentially, to detect the suspicious behaviour of the trolls in spreading negative tweets in some topics and positive in some others.

Despite the better performance of our approaches, there are still a couple of limitations. Our approaches consider that we have a general knowledge of the issues that trolls address. As we showed in Section 4.1, we extracted seven topics that are used by trolls to seed discord. The number of topics is not automatically set, and supervision by human knowledge is needed. For instance, a topic different than the US 2016 elections needs a different number of topics. Another aspect is that our feature set is language-dependent. Recently, during the Italian 2019 elections for the European Parliament, some journalists claimed that they noticed a rise in the amount of fake Twitter accounts that tweeted to affect the public decisions¹⁴. To apply for instance our approaches successfully on an Italian corpus, we would need Italian language lexicons (emotions, stance cues, etc.) and an Italian POS tagger.

9. Conclusion

In this paper, we present two text-based approaches to detect social media trolls, namely IRA trolls. Due to the anonymity characteristic that social media provide to users, these kinds of suspicious behavioural accounts have started to appear. We built machine learning models based on topic and profiling features that in a cross-validation evaluation achieved F1 values of 0.88 and 0.91. We applied a topic modeling algorithm to go behind the superficial textual information of the tweets. Our experiments showed that the extracted topics boosted the performance of the proposed models when coupled with other affective and lexical 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, linguistic, and thematic perspectives. Through the manually checking of IRA accounts, we noticed that frequently irony was employed. As a future work, it would be interesting to try to identify these accounts by integrating an irony detection module, although irony detection is still an open research topic and results may be far from accurate.

Acknowledgments

This work is partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the program “Investissements d’Avenir” (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris - ANR-18-IDEX-0001. The work of

¹⁴<https://www.thelocal.it/20180802/russian-troll-factory-tweets-attempted-influence-italian-elections>

Paolo Rosso was partially funded by the Spanish MICINN under the research project MIS-FAKEHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

- [1] J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, E. Gilbert, Still out there: Modeling and Identifying Russian Troll Accounts on Twitter, arXiv preprint arXiv:1901.11162 (2019).
- [2] E. M. Clark, J. R. Williams, C. A. Jones, R. A. Galbraith, C. M. Danforth, P. S. Dodds, Sifting Robotic from Organic Text: A Natural Language Approach for Detecting Automation on Twitter, *Journal of Computational Science* 16 (2016) 1–7.
- [3] R. L. Boyd, A. Spangher, A. Fourney, B. Nushi, G. Ranade, J. Pennebaker, E. Horvitz, Characterizing the Internet Research Agency’s Social Media Operations During the 2016 US Presidential Election using Linguistic Analyses, *PsyArXiv* (2018).
- [4] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, J. Blackburn, Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and their Influence on the Web, in: *Companion Proceedings of The 2019 World Wide Web Conference*, ACM, 2019, pp. 218–226.
- [5] G. Gorrell, M. E. Bakir, I. Roberts, M. A. Greenwood, B. Iavarone, K. Bontcheva, Partisanship, Propaganda and Post-Truth Politics: Quantifying Impact in Online, arXiv preprint arXiv:1902.01752 (2019).
- [6] A. Badawy, K. Lerman, E. Ferrara, Who Falls for Online Political Manipulation?, in: *Companion Proceedings of The 2019 World Wide Web Conference*, ACM, 2019, pp. 162–168.
- [7] A. Badawy, E. Ferrara, K. Lerman, Analyzing the Digital Traces of Political Manipulation: the 2016 Russian Interference Twitter Campaign, in: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2018, pp. 258–265.
- [8] K. Lee, B. D. Eoff, J. Caverlee, Seven months with the devils: A long-term study of content polluters on twitter, in: *Fifth international AAAI conference on weblogs and social media*, 2011.
- [9] F. Rangel, P. Rosso, Overview of the 7th author profiling task at pan 2019: Bots and gender profiling, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), *CLEF 2019 Labs and Workshops, Notebook Papers*, CEUR-WS.org, 2019.
- [10] F. Johansson, Supervised classification of twitter accounts based on textual content of tweets, in: *CLEF 2019 Labs and Workshops, Notebook Papers*, volume 2019, CEUR-WS.org, 2019.
- [11] J. P. Dickerson, V. Kagan, V. Subrahmanian, Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?, in: *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, IEEE, 2014, pp. 620–627.
- [12] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, F. Menczer, Botornot: A system to evaluate social bots, in: *Proceedings of the 25th International Conference Companion on World*

- Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 273–274.
- [13] A. Ng, This was the Most Viewed Facebook ad Bought by Russian Trolls, 2018. URL: <https://www.cnet.com/news/this-was-the-most-viewed-facebook-ad-bought-by-russian-trolls/>.
- [14] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, *Journal of machine Learning research* 3 (2003) 993–1022.
- [15] S. M. Mohammad, P. D. Turney, Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon, in: *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, Association for Computational Linguistics, 2010, pp. 26–34.
- [16] S. Frenda, B. Ghanem, M. Montes-y Gómez, Exploration of Misogyny in Spanish and English Tweets, in: *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150, CEUR-WS, 2018, pp. 260–267.
- [17] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, Semeval-2016 Task 6: Detecting Stance in Tweets, in: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 31–41.
- [18] H. Bahuleyan, O. Vechtomova, UWaterloo at SemEval-2017 Task 8: Detecting Stance Towards Rumours with Topic Independent Features, in: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 461–464.
- [19] B. Ghanem, A. T. Cignarella, C. Bosco, P. Rosso, F. M. R. Pardo, UPV-28-UNITO at SemEval-2019 Task 7: Exploiting Post’s Nesting and Syntax Information for Rumor Stance Classification, in: *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 1125–1131.
- [20] J. B. Hooper, On Assertive Predicates, volume 4, In J. Kimball, editor, *Syntax and Semantics*, 1974.
- [21] M. Recasens, C. Danescu-Niculescu-Mizil, D. Jurafsky, Linguistic Models for Analyzing and Detecting Biased Language, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, 2013, pp. 1650–1659.
- [22] P. Kiparsky, C. Kiparsky, *Fact*, Linguistics Club, Indiana University, 1968.
- [23] L. Karttunen, Implicative Verbs, *Language* (1971) 340–358.
- [24] S. Volkova, S. Ranshous, L. Phillips, Predicting Foreign Language Usage from English-Only Social Media Posts, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, 2018, pp. 608–614.
- [25] Y. R. Tausczik, J. W. Pennebaker, The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods, *Journal of language and social psychology* 29 (2010) 24–54.
- [26] J. Graham, J. Haidt, B. A. Nosek, Liberals and Conservatives Rely on Different Sets of Moral Foundations, *Journal of personality and social psychology* 96 (2009) 1029.
- [27] S. Malmasi, K. Evanini, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, Y. Qian, A Report on the 2017 Native Language Identification Shared Task, in: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 2017, pp. 62–75.
- [28] A. Cimino, F. Dell’Orletta, Stacked Sentence-Document Classifier Approach for Improving

- Native Language Identification, in: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 430–437.
- [29] I. Markov, L. Chen, C. Strapparava, G. Sidorov, CIC-FBK Approach to Native Language Identification, in: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 374–381.
- [30] C. Goutte, S. Léger, Exploring Optimal Voting in Native Language Identification, in: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 367–373.
- [31] R. Zheng, J. Li, H. Chen, Z. Huang, A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques, *Journal of the American society for information science and technology* 57 (2006) 378–393.
- [32] M. Bhargava, P. Mehndiratta, K. Asawa, Stylometric Analysis for Authorship Attribution on Twitter, in: International Conference on Big Data Analytics, Springer, 2013, pp. 37–47.
- [33] M. Sultana, P. Polash, M. Gavrilova, Authorship Recognition of Tweets: A Comparison Between Social Behavior and Linguistic Profiles, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2017, pp. 471–476.
- [34] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical Attention Networks for Document Classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.
- [35] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation* 9 (1997) 1735–1780.
- [36] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
- [37] B. C. Boatwright, D. L. Linvill, P. L. Warren, Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building, *Resource Centre on Media Freedom in Europe* (2018).
- [38] W. Haskell, People Explaining their 'Personal Paradise' is the Latest Hashtag to Explode on Twitter, 2015. URL: <https://www.businessinsider.com.au/hashtag-games-on-twitter-2015-6>.
- [39] B. Dhingra, Z. Zhou, D. Fitzpatrick, M. Muehl, W. W. Cohen, Tweet2Vec: Character-Based Distributed Representations for Social Media, in: The 54th Annual Meeting of the Association for Computational Linguistics, 2016, p. 269.