Fake News Spreaders Profiling Through Behavioural Analysis

Notebook for PAN at CLEF 2020

Matteo Cardaioli\textsuperscript{1,2}, Stefano Cecconello\textsuperscript{1}, Mauro Conti\textsuperscript{1}, Luca Pajola\textsuperscript{1}, and Federico Turrin\textsuperscript{1}

\textsuperscript{1} Department of Mathematics, University of Padua, Padua
\textsuperscript{2} GFT Italy, Milano, Italy
\{matteo.cardaioli, stefano.cecconello, conti, pajola, turrin\}@math.unipd.it

Abstract The growth of social media and the people interconnection led to the digitalization of communication. Nowadays the most influential politicians or scientific communicators use the media to disseminate news or decisions. However, such communications media can be used maliciously to spread the so-called fake-news in order to polarise public opinion or to deny scientific theories. It is therefore important to develop intelligent and accurate techniques in order to identify the spreading of fake-news. In this paper, we describes the methodology regarding our participation in the PAN@CLEF Profiling Fake News Spreaders on Twitter competition. We propose a supervised Machine-Learning (ML) based framework to profile fake-news spreaders. Our method relies on the combination of Big Five personality and stylometric features. Finally, we evaluate our framework detection capabilities and performance with different ML models on a tweeter dataset in both English and Spanish languages.

1 Introduction

Social Networks, such as Twitter, Facebook, or Instagram, are nowadays the main source of information for millions of people. The success of Social Networks is mainly due to their usability and their free access to every kind of information [20]. However, Social Networks lack in content control, and this allowed the emergence of several malicious phenomena such as the aggregation of terrorist groups [27] and the spreading of hate messages [2] or fake-news messages [17]. Due to their impact, in recent years, the analysis of online platforms as Online Social Networks (OSNs), blogs, and forums attracted a wide area of security researchers. Different tasks can be found, from bot detection in OSNs [1, 18] to hate speech detection [5, 13].

Among the different Social Network analysis, fake-news detection is increasingly attracting the researcher and industrial attention due to the impact of such a phenomenon. In [11], the authors investigated on the fake-news spreading on Twitter during the 2016 U.S. presidential election showing dramatic results in term of fake-news exposition and
sharing. To limit the spreading of fake news, it is also important to identify in advance users more likely believe in them.

These results highlight the importance of developing accurate techniques to identify and prevent fake-news spreading. In this context, ML plays an important role because it can help to classify fake-news in an automatic manner, that otherwise would be impossible for a human operator. Common approaches for the identification of fake-news rely on traditional methods of features extraction such as: Bag of Words, TF-IDF, stylometry; but also novel methods such as fact-checkers automatic detectors [24].

An unexplored field related to fake-news detection is the identification of fake news spreaders, which is the goal of PAN@CLEF Profiling Fake News Spreaders on Twitter competition [21]. In this work, we investigate how behavioural features (i.e., personality, stylometry) can discriminate a fake-news spreader.

We summarise the contribution of the paper as follows:

- We present and implement a framework for profiling fake-news spreaders by exploiting behavioural features (i.e., personality, stylometry).
- We compare and evaluate the classification performances of different supervised ML models in a Twitter feed dataset.

The remainder of the paper is organised as follows. Section 2 provides an overview of the related work. Sections 3 and 4 outline respectively our methodology we used to identify the fake-news tweets and the evaluation approach we implemented. Finally, Section 5 concludes the paper.

2 Related works

The analysis of phenomena in social networks can be very complex and thus standard Natural Language Processing (NLP) techniques might not be very effective. For profiling tasks [23] (e.g., gender, race) researchers started to look in other disciplines (e.g., psychology) to enrich tasks’ feature space; for example, common trends involve techniques that use “emotions” or “behaviours”.

**Emotion-based profiling.** In online platforms such as social networks, users interact with each other through posts, comments, and messages, sharing ideas, feelings, and emotions. Such features can be used to understand better both cultural-insights and human behaviours. For example, Kusen et al. [19] analyze users’ emotions during different events and found that they tend to conform to the emotional valence of the respective real-world event; however, users also tend to react with positive messages during negative events, exhibiting a strong shifted emotion. Recently, several works show that emotional-based features can improve the efficacy of tasks such as clickbait identification [8] and fake-news detection [9].

**Behavioural-based profiling.** Temporal patterns can be a powerful tool to profile phenomena. For example, behavioural-based techniques are a consolidated standard in biometric [26]. In the context of online platforms security, behavioural features are used to detect bots [3, 14, 16]. Chu et al. [3] use behavioural biometrics such as mouse and
keystroke dynamics to discriminate bots from humans. Similarly, Hall et al. [14] to de-
tect bots activities in Wikipedia developed an algorithm that leverages on the fact that
bots patterns are often distinct from human patterns. In massively multiplayer online
role playing games (MMORPG) bots are often used to obtain games items that can be
exchanged with real money. Kang et al. designed a detection approach to solve such a
problem by identifying behavioural patterns that are unique for bots in MMORPG [16].

3 Methods

This study aims to detect Twitter user profiles that are keen to be spreaders of fake-news.
Our analysis wants to retrieve information that can provide insights into the attitude to
spread fake-news, rather than identifying potential fake-news among the users’ tweets.
Our approach is therefore focused on the extraction of behavioural characteristics of the
user such as personality and writing style.

3.1 Dataset description

The provided dataset contains 60000 tweets (100 per user) derived from 600 distinct
Twitter accounts. For each user profile are also reported: the language, a unique author
ID, and a binary label that defines the class (the information regarding which label cor-
responds to each class is not provided by the organizers due to GDPR reasons). The
users are equally distributed among two languages: 300 English profiles and 300 Span-
ish profiles. Furthermore, the dataset contains an equal number of fake-news spreader
and not fake-news spreader profiles. To guarantee better profile anonymisation, sensible
information contained in the original tweets has been obfuscated using some keywords.
In particular, the following keywords were used: “user”, “rt” (re-tweet), “hashtag”,
“URL”. The keywords are always preceded by a hashtag and are always capitalised.

3.2 Pre-processing and Feature Extraction

In this study, we wanted to use a multidisciplinary approach, with the aim of exploring
the impact of uncommon features on identification of fake news spreaders. We focused
our analysis on two main behavioural characteristics: writing style and personality.

Writing Style Features  Stylometry features have been used to solve several tasks [7].
Guided by their popularity, we decide to include 10 stylometric features that summarise
the writing style of the users. In particular, we evaluated:

– **Diversity score** describes how tweets are novel between each other. We use the de-
  finition given in [25], and it is defined as the average diversity of the user tweets.
  Given a list of tweets $C$, the diversity of the tweet $t_i$ is defined as $1 - \max(Jac(x_i, x_j))$,
  where $i \neq j$ and $j \in [0, ..., |C|]$. “Jac” is the Jaccard similarity of two sets.
– **Readability score** is the average of the “Flesch reading ease” of the tweets. The
  score is calculated with the python library “Textstat”\(^3\). This particular metric is
  defined over a multiple set of languages, such as English and Spanish.

\(^3\) https://pypi.org/project/textstat/
– Hash avg is the average hashtags per tweet.
– Usr avg is the average of mentions per tweet.
– Url avg is the average of URLs per tweet.
– Retweet is the average retweets per tweet.
– Lower is the average lowercase characters per tweet.
– Upper is the average uppercase characters per tweet.
– Punctuation is the average of punctuation characters per tweet.
– Alpha is the average of alphabetical characters per tweet.

**Personality Features** To extract personality features, we firstly performed a pre-processing on the dataset. We removed all the keywords used to obfuscate sensible information (e.g., #URL, #RT). We then merged all the tweets, creating a unique corpus for every user. We used Watson Personality Insights - IBM \(^4\) to retrieve personality information from written text. In particular, giving as input the corpus and the language (i.e., “Spanish” or “English”). The output of IBM Watson consists of a JSON containing: Needs, Values, and Big Five personality characteristics. For each of them, the service provides a percentile score. The higher this score is, the greater is also the presence of the specific personality trait for the user. Needs describe at a high level those aspects of a product that are likely to resonate with the author of the input text. Values describe motivating factors that influence the author’s decision-making. The Big Five model \([15, 22]\), or OCEAN, is one of the most famous in the study of personality. According to this theory, personality can be divided into five independent traits. Below, we provide a brief definition for each trait, as reported in \([12]\).

– **Agreeableness** is a person’s tendency to be compassionate and cooperative toward others.
– **Conscientiousness** is a person’s tendency to act in an organised or thoughtful way.
– **Extraversion** is a person’s tendency to seek stimulation in the company of others.
– **Emotional Range**, also referred to as Neuroticism or Natural Reactions, is the extent to which a person’s emotions are sensitive to the person’s environment.
– **Openness** is the extent to which a person is open to experiencing different activities.

In Figure 2 we compare the average values of the Big Five percentiles for both Class 0 and Class 1. We can notice the following: i) on average, Class 0 tends to have lower Agreeableness and Extraversion values and ii) on average, Class 0 tends to have higher Emotional Range values.

Many studies on the Big Five present a two-level hierarchy, with the five domains discussed above assuming more specific traits, called “facets” at a second level \([6]\). In particular, each Big Five trait can assume six facets as described in \([4]\). The total number of personality features are 54.

Figure 1 depicts a sunburst chart reporting the percentile scores returned by Watson APIs. The chart reports the best score for each personality characteristics (Big Five, Needs, and Values). For Needs and Values, the chart reports the score for the features provided by Watson (respectively 12 and 5 values). For Big Five, a supplementary level reports the percentile for each one of the five characteristics, followed by the same score for each facet.

\(^4\) https://www.ibm.com/watson
4 Evaluation

In this section, we discuss the process implied to build our model. The pipeline of our approach is defined as follows:

- Pre-processing. Special token cleaning (for personality features only).
- Feature extraction. Stylometric and personality features. The final set of features consists of the concatenation of both stylometric and personality features, with the addition of the typing language (for a total of 64 features). This approach allow us to implement cross-language models.
- Feature selection. The dimensionality of data is reduced by applying a KBest feature selection algorithm.
- Model. Simple binary classification models (e.g., Random Forest).
Different models are evaluated to discriminate between *fake-news spreaders* and *non-spreaders*: Decision Tree (DT), K-Nearest Neighbour (KNN), Support Vector Classifier (SVC), and Random Forest (RF). For the model selection, we used a repeated nested-cross-validation (100 times for 80-20 random splits) with 5-fold inner cross-validation, applied to the aforementioned pipeline. This approach is used to assess the robustness of the prediction for every ML model among different splits. According to Table 1 we selected RF classifier that showed the highest average accuracy and the lowest variability in nested-cross-validation. Finally, we trained the RF classifier on the whole dataset, using the same pipeline with a 5-fold cross-validation. To obtain the best RF’s hyperparameters a 5-fold cross-validation is applied on the same parameter grid shown in Table 1, obtaining `n_estimators = 25` and `max_depth = 4`. Finally, RF is trained on the whole dataset, using these hyperparameters.

![Radar chart reporting the average Big Five scores for “Class 0” and “Class 1” classes.](image)

**Figure 2.** Radar chart reporting the average Big Five scores for “Class 0” and “Class 1” classes.

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>Values</th>
<th>Training Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT max_depth</td>
<td>[2, ..., 5]</td>
<td>0.75 ± 0.05</td>
<td>0.65 ± 0.05</td>
</tr>
<tr>
<td>KNN n_neighbors</td>
<td>[3, ..., 10]</td>
<td>0.81 ± 0.03</td>
<td>0.71 ± 0.04</td>
</tr>
<tr>
<td>SVC c</td>
<td>[10^{-2}, 10^{-1}, 1, 10]</td>
<td>0.94 ± 0.08</td>
<td>0.73 ± 0.04</td>
</tr>
<tr>
<td>SVC γ</td>
<td>[10^{-3}, 10^{-2}, 10^{-1}, 1, 10]</td>
<td>0.94 ± 0.08</td>
<td>0.73 ± 0.04</td>
</tr>
<tr>
<td>RF n_estimators</td>
<td>[25, 50, 100, 200]</td>
<td>0.87 ± 0.03</td>
<td>0.74 ± 0.03</td>
</tr>
<tr>
<td>RF max_depth</td>
<td>[3, 4, 5]</td>
<td>0.87 ± 0.03</td>
<td>0.74 ± 0.03</td>
</tr>
</tbody>
</table>

**Table 1.** Training and Test accuracy of the nested validation for the different models.
5 Conclusions

In this work, we describe the approach we used for the PAN@CLEF Profiling Fake News Spreaders on Twitter challenge. Our proposed framework relies on the combination of linguistic features with psychological traits with simple machine learning methods (e.g., Random Forest). Our method reached the following accuracy: 0.6750 (English corpus), 0.7150 (Spanish corpus), and 0.6950 (average between the two corpus). Our work shows the feasibility of using psychological features to determine whether a user is a fake news spreader or not. The intuition on using psychological behaviours in the fake news domain is also confirmed by the recent work proposed by Giachanou et al. [10], where the authors combine linguistic patterns with personality scores to distinguish between fake news spreaders and checkers. We believe that future research directions could benefit from this, enlarging the feature space of the data.

References


5 This paper is published after the time of our submission.


