Profiling fake news spreaders through stylometry and lexical features. UniOR NLP @PAN2020

Notebook for PAN at CLEF 2020

Raffaele Manna[1], Antonio Pascucci[1], and Johanna Monti[1]

[1]"L' Orientale" University of Naples - UNIOR NLP Research Group
{rmanna, apascucci, jmonti}@unior.it

Abstract In this paper, we describe our approach to address the *Profiling Fake News Spreaders on Twitter* task at PAN 2020[1]. The aim of the task is to profile users who are used to spread (consciously or unconsciously) fake news in two languages, namely English and Spanish. We use different machine learning algorithms combined with strictly stylometric features, categories of emojis and a bunch of lexical features related to the fake news headlines vocabulary. As results of the final official runs, our models achieve an accuracy of 72.50% for the Spanish sub-task (using the Logistic Regression algorithm) and an accuracy of 59.50% for the English sub-task (using the Random Forest algorithm).

1 Introduction

The flow of information and news is growing day by day on social media. Social media platforms now represent the primary means for personal information on events and facts of different nature that happen around us in the real world. It could be said that social media are what the *agorà* was once to the ancient Greeks, namely a crowded place where people meet and exchange opinions and information on everyday events. It also means that news are often not credible because they can be shared by unreliable sources. In fact, these types of news show manipulative content and expose an alternative of the facts. In other words, news does not represent reality and tries to influence the reader [14]. Furthermore, the massive diffusion of fake news involves the polarization of public opinion on certain debated issues, often increasing offensive attitudes and hate speech towards other points of view and other groups of people [20]. In this context, cybersecurity techniques, digital forensics investigations and computational stylometry are essential in monitoring and identifying the main sources of fake news. Moreover, a second application scenario to counter the spread of fake news is the task of profiling users based on their susceptibility in sharing texts with inaccurate information.

The 2020 edition of the PAN Author Profiling task[2] focuses on the classification of

[1] https://pan.webis.de/clef20/pan20-web/index.html

[2] https://pan.webis.de/clef20/pan20-web/author-profiling.html

potential fake news spreaders on Twitter, whether they propagate fake news intentionally or unintentionally.

In this paper, we propose a machine learning classification approach based on stylometric features along with two lexical category features: persuasive words associated with fake news [19] and words associated with subjectivity. The paper is organized as follows: in section 2 we present related work, in section 3 we present the problem by describing the author profiling task proposed at PAN 2020 and the dataset provided by the shared task organizers and then we focus on the features and the algorithms used. In section 4, we show the results obtained by our models and the evaluation framework TIRA [15]. Finally, in section 5 we outline the conclusions.

2 Related Work

Considering the massive creation and rapid spread of fake news and the potential threat to the opinion of users, in recent years particular attention was paid to social media and how these represent breeding grounds for tendentious contents spreading, partisan articles dissemination and, in general, invented or modified news to achieve particular purposes.

Scholars have shown how the impact of fake news can affect the creation of the so-called "echo-chambers" as well as can influence the opinions of users during the months before the 2016 US presidential election [1]. Besides, other scholars have shown the role of bots in the diffusion of fake news and misinformation on particular political events to damage a politician [2, 18]. Starting from the categorization of concepts related to fake news, Zhou and Zafarani focused on the different aspects of fake and then analyzed the false information conveyed up to the role played by users [23]. Potthast et al. focused on the hyper-partisan news writing style linked to fake news [16]. The results show how left-wing and right-wing writing styles are very similar and easily distinguishable from the mainstream news. The same research reported some difficulties in the detection of fake news based only on style features.

Scholars have also investigated the role played by users in spreading fake news on social media. Often users mix their personal contents with fake news for either satirical or malicious purposes making the monitoring and classification of content and profiles controversial. In fact, Ghanem et al. attempted to identify Twitter accounts suspected of spreading fake news. The approach is based on a Recurrent Neural Network (RNN) along with semantic and stylistic features [5]. The *CheckerOrSpreader* [6] is a system based on a Convolutional Neural Network (CNN) and aims to differentiate between checkers and spreaders. It consists of two different components: word embedding component (based on the tweets posted on the users' timeline), and psycho-linguistic component that represents style pattern and personality traits that derive from the textual content. Shu, Wang and Liu [21] focused on the correlation between user profiles and fake/real news, showing that there are specific users that are most likely to trust fake news. These users reveal different features from the users that are most likely to trust real news.

3 Dataset

PAN event takes its name from the *International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN)* [22] held in 2007. As the years passed, PAN has become the main event for computational stylometry scholars. PAN event can be described as a series of scientific events and shared task on issues relating to digital forensics and computational stylometry, such as authorship analysis (profiling and identification), computational ethics, and plagiarism detection.

In this edition, four different shared task have been presented: *Authorship Verification*, *Celebrity Profiling*, *Profiling Fake News Spreaders on Twitter* and *Style Change Detection*[3].

Our team (UniOR NLP) decided to take part in the *Profiling Fake News Spreaders on Twitter* task. The aim is to build a model able to identify possible fake news spreaders on social media in a multilingual perspective: data are in fact in Spanish and English. The dataset is made up of Twitter accounts for both languages considered in the task (i.e. Spanish and English). Each account is composed of the author feed of 100 concatenated tweets.

Languages	Train	Test	Total
English	300	200	500
Spanish	300	200	500

Table 1. Number of Twitter users in the dataset.

As shown in Table 1, the dataset is divided into two sets, train and test, for a total of 500 Twitter accounts taken for the construction of the dataset task.

The train set (made available for download by the task organizers) consists of 300 xml files per language, each one containing the author feed and named with an alphanumeric code relating to the identity of the author. Moreover, URLs, mentions and hashtags have been replaced with generic tags for tweets contained in the author feed. The train set is balanced between the two classes. In fact, among 300 xml files, it contains 150 accounts belonging to the spreaders class and 150 accounts belonging to the no-spreaders class.

4 Methodology

In order to identify and classify fake news spreaders, we include in our models two categories of features. The first category is related specifically to stylometric features. The second one focuses on lexical features divided into i) lexical elements expressing personal opinion in online communications and ii) clickbait verbs and expressions in fake news headlines [12, 13].

[3] https://pan.webis.de/clef20/pan20-web/index.html

For these two groups of features, we used a bunch of features recognized as crucial to identify fake news [8]. The features computed by our model for both languages are listed and described below. For each Twitter account we used the following features:

- **Emoji**: We calculated the average number of emojis for each account divided by the total number of emojis for each class. In addition, we added the average number of emojis belonging to several emotional characteristics and different characters represented by emoji. We considered the emojis contained in the Unicode Emoji List[4]. From this list, we selected and used emoji characters related to face-affection, face-tongue, face-neutral-skeptical, face-hand, face-concerned, emotions and country-flag.
- **Stylometric Features**: The average number of each stylistic features of the tweets divided by the total number of each stylistic features for each class. These characteristics are: URLs count; space count; words count; initial capital letter words count; capital words count; digits count; punctuation marks count; operators count; average text length; brackets count; question and exclamation marks count; slashes count; retweet, hashtag and user tags count; quotes style count and ellipsis count.
- **Lexical features**: We designed and computed the average number of the presence of a series of lexical items, in both languages, related to:

  1. Groups of words expressing personal opinions in addition to personal pronouns;
  2. Verbs and expressions related to clickbait headlines.

  As an example respectively for the two categories, groups of words, words and typing shortcut in online communication such as: 1) "mine", "myself", "I", "IMO", "IMHO", "yo", "tu", "personalmente" among others; 2) "videos", "link", "directa", "latest", "click", "últimas", "última hora" among others were used.

As a first step, machine learning algorithms combined with stylometric features, categories of emojis and a bunch of lexical features have been tested in order to detect the most performing model. We decided to run different machine learning algorithms fed with the selected features into the virtual machine assigned to us by the organizers. Then, we chose the best performing algorithm for each language on the basis of the results obtained on the training set.

During the development phase, we used well-known classifiers [10], namely Logistic Regression (LR) [9], Random Forest (RF) [11], Multinomial Naïve Bayes (MNB) [7], Support Vector Machine (SVM) [3] and Gradient Boosting classifer (GBC) [4]. All these machine learning classifiers are provided by the Python Scikit-learn library[5]. We decided to keep the basic classifier hyper-parameter in order to evalute the models only on the basis of stylometric and lexical features.

The submitted version of our model first classifies and predicts Twitter accounts in English, then classifies and predicts the Spanish ones.

[4] https://unicode.org/emoji/charts/full-emoji-list.html
[5] https://scikit-learn.org/stable/

In order to evaluate our selected classifiers, we created our own test set, splitting the train set into 70% training data and 30% test data[6].

Languages	LR	RF	MNB	SVM	GBC
English	0.57	**0.64**	0.52	0.51	0.57
Spanish	**0.81**	0.80	0.73	0.73	0.72

Table 2. Accuracy results obtained by the algorithms selected against our test set.

As shown in Table 2, the best performing algorithms are Random Forest for English and Logistic Regression for Spanish. For each algorithm and for both languages, we used the same set of features listed in subsection 3.3.

5 Results

For the final run on blind test set, we set up a model based on the Logistic Regression [9] algorithm for the Spanish sub-task and a model based on the Random Forest [11] algorithm for the English sub-task. To complete the submission of our software, we run our model on the TIRA [15] platform.

Test set	Results
English	59.50%

Table 3. Results in terms of accuracy obtained by our model on the English test set. The model is based on the Random Forest algorithm.

Test set	Results
Spanish	72.50%

Table 4. Results in terms of accuracy obtained by our model on the Spanish test set. The model is based on the Logistic Regression algorithm.

As shown in Tables 3 and 4, our model seems to better profile and predict by far accounts related to Spanish users than Twitter accounts of English users. In addition, we observe that all the features used to profile users seem to be more present in Spanish tweets as they better discriminate the textual style of fake news spreaders accounts in Twitter.

[6] https:scikit-learn.orgstablemodulesgeneratedsklearn.model_selection.train_test_split.html

6 Conclusions

In this paper, we have shown the results achieved by the UniOR NLP team for the *Profiling fake news spreaders task* [17] at PAN 2020. Our approach is based on stylometric features and two lexical category features: clickbait expressions associated with fake news and words expressing personal opinions along with personal pronouns. Our model achieved much better results in the Spanish sub-task (72.50%) compared to those of the English sub-task (59.50%).

Acknowledgements

This research has been carried out in the context of two innovative industrial PhD projects in computational stylometry supported by the PON Ricerca e Innovazione 2014-20 and the POR Campania FSE 2014-2020 funds.

We sincerely thank the PAN organizers for the work done in order to enable us to submit our system.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of economic perspectives **31**(2), 211–36 (2017)
2. Bessi, A., Ferrara, E.: Social bots distort the 2016 us presidential election online discussion. First Monday **21**(11-7) (2016)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
4. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
5. Ghanem, B., Ponzetto, S.P., Rosso, P.: Factweet: profiling fake news twitter accounts. arXiv preprint arXiv:1910.06592 (2019)
6. Giachanou, A., Ríssola, E.A., Ghanem, B., Crestani, F., Rosso, P.: The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers. In: International Conference on Applications of Natural Language to Information Systems. pp. 181–192. Springer (2020)
7. Granik, M., Mesyura, V.: Fake news detection using naive bayes classifier. In: 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). pp. 900–903. IEEE (2017)
8. Horne, B.D., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: Eleventh International AAAI Conference on Web and Social Media (2017)
9. Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied logistic regression, vol. 398. John Wiley & Sons (2013)
10. Jain, A., Kasbe, A.: Fake news detection. In: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). pp. 1–5. IEEE (2018)
11. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002), https://CRAN.R-project.org/doc/Rnews/
12. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. arXiv preprint arXiv:1708.07104 (2017)

13. Piotrkowicz, A., Dimitrova, V., Otterbacher, J., Markert, K.: The impact of news values and linguistic style on the popularity of headlines on twitter and facebook. In: Eleventh International AAAI Conference on Web and Social Media (2017)
14. Popat, K., Mukherjee, S., Yates, A., Weikum, G.: Declare: Debunking fake news and false claims using evidence-aware deep learning. arXiv preprint arXiv:1809.06416 (2018)
15. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World. Springer (Sep 2019)
16. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B.: A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638 (2017)
17. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (Sep 2020), CEUR-WS.org
18. Rangel, F., Rosso, P.: Overview of the 7th author profiling task at pan 2019: Bots and gender profiling in twitter. In: Proceedings of the CEUR Workshop, Lugano, Switzerland. pp. 1–36 (2019)
19. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 conference on empirical methods in natural language processing. pp. 2931–2937 (2017)
20. Rosso, P.: Profiling bots, fake news spreaders and haters. In: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language (2020)
21. Shu, K., Wang, S., Liu, H.: Understanding user profiles on social media for fake news detection. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 430–435. IEEE (2018)
22. Stein, B., Koppel, M., Stamatatos, E. (eds.): SIGIR 07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07). CEUR-WS.org (2007), http://ceur-ws.org/Vol-276
23. Zhou, X., Zafarani, R.: Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315 (2018)