Profiling Fake News Spreaders on Twitter based on TFIDF Features and Morphological Process

Notebook for PAN at CLEF 2020

Mohamed Lichouri, Mourad Abbas, and Besma Benaziz

Computational Linguistics Department, CRSTDLA, Algiers. Algeria
{m.lichouri,m.abbas,b.benaziz}@crstdla.dz

Abstract. In this paper, we present a description of our experiments on Profiling Fake News Spreaders on Twitter based on TFIDF Features and Morphological Processes as stemming, lemmatization and part of speech tagging. A comparison study between a set of classifiers has been carried out. The best results were achieved using the model LSVC which yielded an f1-score of 76% and 58.50% for Spanish and English, respectively.

1 Introduction

Fast posting, quick access and free publishing of news in social media is a good motivation to spread news in various fields. However, spreading of news in social media is a double-edged sword because it can be used either for beneficial purposes or for bad purposes (fake news).
According to [21], false information is categorized into eight types: fabricated, propaganda, conspiracy theories, hoaxes, biased or one-sided, rumors, clickbait,and satire news. Twitter has recently detected a campaign[20]organized by agencies from two different countries to affect the results of the last U.S. presidential elections of 2016. Social media allows users to hide their real profiles, which gives them a safe space to spread whatever comes to mind.
The ability to know the features of social media users is a growing field of interest called author profiling. There are three main types of fake news contributors: social bots, trolls, and cyborg users [19]. The social bot is an automatic social media account managed by an algorithm, designed to create posts without human intervention [4]. For example,"studies show that social bots distorted the 2016 US presidential election discussions on a large scale, and around 19 million bot accounts tweeted in support of either Trump or Clinton in the week leading up to the election day" [19]. Similarly, according to Marc Jones and Alexei Abrahams [7], a plague of Twitter bots is roiling the Middle East[20].

The troll is a user of another kind that spreads false news among societies across the Internet. It is a type of user who aims to disrupt online communities

and provoke consumers into an emotional response [3], [19]. For instance, there has been evidence that claims "1,000 Russian trolls were paid to spread fake news on Hilary Clinton," which reveals how actual people are performing information manipulation in order to change the views of others [19]. The troll differs from the bot program because the troll is a real user, while the bot software is automatic. The mixture between the bots and trolls, can produce a type which is not less dangerous than the above. Intelligence in this type lies in the account registered by real users, but use programs to perform activities in social media. With the possibility of switching between the two [19].
In this paper, we are interested in profiling fake news spreaders on Twitter [13] for two languages: English and Spanish using a machine learning model[8].

The paper is organized as follows: in section 2, we present the main works related to profiling fake news spreaders. In section 3, we describe the dataset used in our experiments as well the preprocessing steps that we followed. Our system architecture including feature extraction and classification models is presented in section 4. We summarize the achieved experiments and the results in section 5 and we conclude in section 6.

2 Related work

Author profiling is a problem of growing importance, as it it can be helpful for combating fake news. Indeed, it allows us to differentiate between real and imaginary users, or even to reach everyone who posted fake news. Many works are interested in studying the possibility of obtaining age and gender through formal texts [6], [2]. The writer's age and gender can appear through his publications, including ideas and diversity in linguistic characteristics. In [11], [9], the authors found out that women, at least for English language, use the first single person more than men who use more determinants because they talk about tangible things. This allowed the authors to build the LIWC (Linguistic Inquiry and Word Count), which is effective in author profiling. In [18], a study of (71,000) blogs showed that the linguistic features in blogs are related to age and gender. They got an accuracy of about 80% to determine gender and about 75% to determine age. Author profiling tasks have been organized many years at PAN[1]. Indeed, in [14], the authors describe a large corpus, collected on social networks, and its characteristics, to solve the problem of identifying age and sex. Rangel et. al. [16] continued to focus on aspects of age and gender, where the aim of this work was to analyse the adaptability of the detection approaches when given different genres. For this purpose, a corpus with four different parts (sub-corpora) has been compiled: social media, Twitter, blogs, and hotel reviews. In [15], two new languages have been added, Italian and Dutch, besides a new subtask on personality recognition, to enrich the results obtained previously. In [17], the objective was to predict age and gender from a cross-genre perspective. For this purpose a corpus from Twitter has been provided for training, and different corpora from social media, blogs, essays, and reviews have been provided for

[1] http://webis.de

evaluation. In [7], the objective was to address gender and language variety identification. For this purpose a corpus from Twitter has been provided for four different languages: Arabic, English, Portuguese, and Spanish.

In [5], the authors provide an emotionally infused deep learning network that uses emotional features to identify false information in Twitter and news articles sources. They compared the language of false news to the one of real news from an emotional perspective, considering a set of false information types (propaganda, hoax, click-bait, and satire) from social media and online news article sources. The results show that the detection of suspicious news in Twitter is harder than detecting it in news articles.

3 Dataset and Preprocessing

The dataset is saved and organized as XML files. It is composed of thousands of tweets of several authors (Twitter users). In fact, 500 XML files (corresponding to 500 authors) are provided for English and the same number is reserved for Spanish. Each file includes 100 tweets, which means that the total number of tweets for both English and Spanish is 100.000 tweets written by 1000 authors. Each XML file is coded with an alpha-numeric author-ID and tagged with two labels: 0 or 1. We performed a basic and necessary text preprocessing step which is punctuation and emojis removal. We summarize in table 1 some statistics about the training set for both English and Spanish. We illustrate in figure 1 the different steps of our proposed system which includes preprocessing, features extraction and model training.

	English	Spanish
# authors (XML files)	300	300
# sentences per author (XML file)	30,000	30,000
# words per author (XML file)	717,596	786,965
Max # word per author (XML file)	3,636	5,373
Min # word per author (XML file)	1,524	1,603
Max # char per author (XML file)	12,962	23,588
Min # char per author (XML file)	5,238	5,799

Table 1. PAN Train set statistics for both English and Spanish

4 System architecture

There are four processes that we used in our approach. The input texts are first subject to the first step: stop words removal. After that, we apply the three additional morphological processes which are: stemming, lemmatization and part of speech tagging. After many trials of combinations between these processes, we

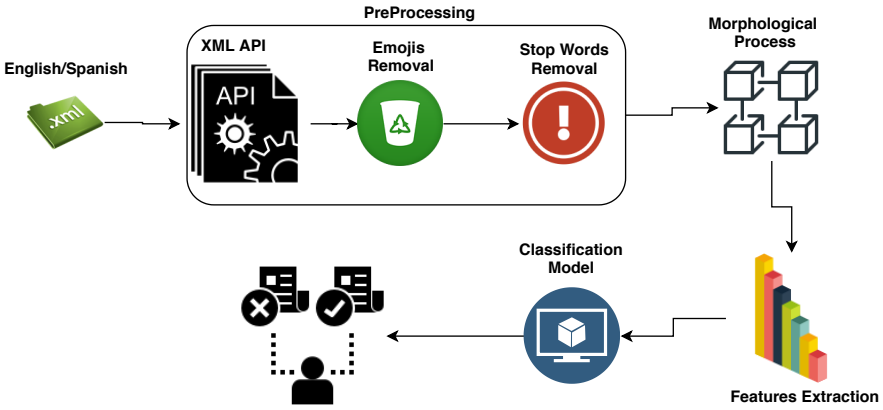

Fig. 1. The architecture of the proposed system.

found out that the combination that gives the best performance is the one resulted in concatenating the text outputs of the three aforementioned processes, in addition to stop words removal, in a single text array. Inspired from [1] in which a union of TFIDF features has given better results for text classification, we have chosen the union of three TF-IDF features (word 5-grams, char 5-grams, char with boundary 5-grams). In addition, we used three classifiers, namely: Linear Support Vector Classification (LSVC), linear model with Stochastic Gradient Descent (SGD) and Ridge Classifier (RDG) [10]. We used the default configuration for selecting the parameters used for each of the aforementioned classifiers.

5 Experiments and results

In order to validate our approach, we split the training data into two sets, 80% for training (240 documents) and 20% for test (60 documents). We tried different classifiers, Linear SVC, SGD and RDG. Results are presented in Table 2 for English and Spanish datasets, in which it is clearly shown that linear SVC and SGD outperformed RDG classifier.

Dataset	Model	F1-score (%)
3*English	LSVC	100
	RDG	99.58
	SGD	100
3*Spanish	LSVC	100
	RDG	99.16
	SGD	100

Table 2. Results using a subset of the training set.

In the final submission, model is trained on the whole training set using LSVC classifier and tested on the official PAN 2020 test set for the author profiling task, on the TIRA platform [12]. Results of final submission are shown in Table 3.

Dataset	Model	F1-score (%)
3*English	LSVC	58.50
	RDG	**61.50**
	SGD	52.00
3*Spanish	LSVC	**76.00**
	RDG	74.50
	SGD	54.50

Table 3. Results of the final submission using LSVC (ranked system), RDG and SGD.

By comparing the results in table 2 and table 3, we notice clearly that the LSVC model performance dropped by 41.42% and 24% for English and Spanish respectively. The RDG classifier is more or less efficient since the recorded score for Spanish was 76.00% and for English 61.50%. The reason behind these results is likely the lack of data.

6 Conclusion

We presented in this paper our approach for identifying authors that tend to spread fake news. We carried out many experiments that led us to select the best features, composed of a union of three TF-IDF features (word 5-grams, char 5-grams and char_wb 5-grams), in addition to three important morphological features: stemming, lemmatization and part of speech tagging. Our system achieved an F1-score of 76% for Spanish and 58.50% for English, which can be improved by increasing the size of the training dataset.

References

1. Abbas, M., Lichouri, M., Freihat, A.A.: St madar 2019 shared task: Arabic fine-grained dialect identification. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop. pp. 269–273 (2019)
2. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 1301–1309 (2011)
3. Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., Leskovec, J.: Anyone can become a troll: Causes of trolling behavior in online discussions. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. pp. 1217–1230 (2017)

4. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM **59**(7), 96–104 (2016)
5. Ghanem, B., Rosso, P., Rangel, F.: An emotional analysis of false information in social media and news articles. ACM Transactions on Internet Technology (TOIT) **20**(2), 1–18 (2020)
6. Holmes, J., Meyerhoff, M.: The handbook of language and gender, vol. 25. John Wiley & Sons (2008)
7. Jones, M.O.: The gulf information war— propaganda, fake news, and fake trends: The weaponization of twitter bots in the gulf crisis. International journal of communication **13**, 27 (2019)
8. Lichouri, M., Abbas, M., Freihat, A.A., Megtouf, D.E.H.: Word-level vs sentence-level language identification: Application to algerian and arabic dialects. Procedia Computer Science **142**, 246–253 (2018)
9. Nerbonne, J.: The secret life of pronouns. what our words say about us. Literary and Linguistic Computing **29**(1), 139–142 (2014)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
11. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology **54**(1), 547–577 (2003)
12. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: Tira integrated research architecture. In: Information Retrieval Evaluation in a Changing World, pp. 123–160. Springer (2019)
13. 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`
14. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation. pp. 352–365. CELCT (2013)
15. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF. p. 2015. sn (2015)
16. Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W., et al.: Overview of the 2nd author profiling task at pan 2014. In: CEUR Workshop Proceedings. vol. 1180, pp. 898–927. CEUR Workshop Proceedings (2014)
17. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF **2016**, 750–784 (2016)
18. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs. vol. 6, pp. 199–205 (2006)
19. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter **19**(1), 22–36 (2017)
20. Vijaya Gadde and Yoel Roth: Enabling further research of information operations on Twitter. `https://blog.twitter.com/en\_us/topics/company/2018/`

enabling-further-research-of-information-operations-on-twitter.html (2018), online; accessed 25 Juillet 2020
21. Zannettou, S., Sirivianos, M., Blackburn, J., Kourtellis, N.: The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. Journal of Data and Information Quality (JDIQ) **11**(3), 1–37 (2019)