Sadness and Fear: Classification of Fake News Spreaders' Content on Twitter

Notebook for PAN at CLEF 2020

Irene Russo

ILC CNR, Italy
irene.russo@ilc.cnr.it

Abstract The vast amount of accurate and inaccurate information circulating on the internet requires computational methodologies to detect low-quality content. This kind of content often constitutes *fake news*, as in the PAN @ CLEF 2020 competition *Profiling Fake News Spreaders on Twitter*. This competition asks for systems that identify possible fake news spreaders on social media as a first step to prevent fake news from being propagated among online users.
In this paper, the methodology used for this classification task is reported. Preprocessing of the data and the features extracted to classify fake news spreaders is explained. A regression-as-classification approach that enables the representation of being a fake news spreader as a gradable one is proposed. The performance (accuracy) on the training and the test set with the different sets of features is reported.

1 Introduction

Nowadays, news production is not exclusive to official media outlets: everybody can report about events. This tendency has positive consequences on the freedom of speech - especially in countries where this fundamental human right is menaced - but it also presents several risks. The vast amount of accurate and inaccurate information circulating on the internet requires computational methodologies to detect low-quality content. The fake information that spreads on social media can be dangerous for public debates on societal issues, increasing the general level of anxiety and affecting the behavior of the population in case of emergency [1]. User-generated content such as pictures and short videos are a potential source of rumors that should be carefully verified [5]. Similarly, reporting news with link sharing can be harmful, especially if the news source is not reliable [4].
On social media information, disinformation (intentionally false content, created to cause harm) and misinformation (false content shared without the user realized it) coexist [11], making it hard to detect reliable channels of information.
We can dedicate a limited amount of time and attention to the verification of a source of information; moreover, repetitions of rumors make them more plausible for everyone

[8].
Harmful content is often labeled as *fake news*, as in the PAN @ CLEF 2020 competition *Profiling Fake News Spreaders on Twitter*[7]. This competition asks for systems that identify possible fake news spreaders on social media as a first step to prevent fake news from being propagated among online users.
Fake news is a standard label used in the NLP community, a trendy term denoting, in reality, many different phenomena that require various features/approaches to be detected and classified: rumors, propaganda, satire, hoaxes, etc. The ubiquity of the term hides the fact that the NLP community lacks an informed typology of fake news types. This typology would need insights from political science and cognitive psychology to discover the most harmful kind of fake news: someone spreading inaccurate information about a pandemic is not dangerous as someone tweeting about last gossip involving Jennifer Lopez.
Fake news are not directly the focus of PAN @ CLEF 2020 competition *Profiling Fake News Spreaders on Twitter* [7]. The organizers are instead concerned with the identification of Twitter profiles that frequently share articles with inaccurate information (intentionally or not), contributing to the creation and propagation of fake news online. According to the organizers, the identifying possible fake news spreaders on social media is the first step towards preventing fake news from being propagated among online users.
In this paper, the experiments aiming at this classification task are reported. Pre-processing of the data and the features extracted to classify fake news spreaders is explained. A regression-as-classification approach that enables the representation of being a fake news spreader as a gradable one is proposed. The performance (accuracy) on the training and the test set with the different sets of features is reported.

2 Related Works

The starting hypothesis of the methodology reported in this paper is that emotions have a crucial role in identifying fake news spreaders because fake content contains emotionally charged words. The role of emotions and emotions' intensity for the detection of fake news has been investigated by [3] that propose EmoCred, a system incorporating emotional signals into a LSTM neural network for the classification of credible and non-credible claims contained in fact-checking datasets. They experimented with lexicon-based emotional analysis, the emotional intensity of words, and a neural network for generating the intensity level of emotional reactions, achieving an accuracy ranging from 0.608 to 0.628 (depending on the dataset and the methodology tested). The results improve the baseline - a LSTM for classification of texts - showing the relevance of emotional signals for fake news classification.
With a focus on the personality traits of fake news spreaders and fake news checkers on Twitter, the methodology proposed by [**?**] addresses the problem of fake news at the users' level using also linguistics patterns found in users' posts to decide if a user is a potential spreader or checker. Their system - CheckerOrSpreader - is a model based on a CNN network and handcrafted features that refer to the linguistic patterns and personality traits and can classify a user as a potential fake news checker or spreader (0.59

as F1 score).
Linguistic features that characterize fact-checkers on Twitter have been analyzed by [10] to create a deep learning framework that generates responses with fact-checking intention. Fact-checkers prefer a formal language, avoiding swear words and Internet slang; the text generation framework proposed outperforms other text generation approaches quantitatively and qualitatively.
Apart from textual features, user social engagements can be used to distinguish users that share real news from users who share fake news on Twitter [9]. For example, a comparative analysis of explicit and implicit profile features reveals that users sharing fake news tend to express more "favor" actions. Their predicted age is slightly bigger when compared with users that share real news.

3 Methodology

With the assumption that the property of being a fake news spreader could be a gradable one dependent on several characteristics of user-generated content, the classification task proposed by PAN @ CLEF 2020 competition *Profiling Fake News Spreaders on Twitter* was addressed as a regression one. A random forest regressor [2] was implemented because it outperforms other regressions algorithms on this dataset. Since the random forest regressor's output is a decimal number, it has been rounded to get the reliability class for each processed instance and then compute the accuracy.

3.1 Data Pre-processing

The set of features used in the regression-as-classification experiments concern stylistic aspects (e.g., use of emphatic punctuation marks), intended communicative functions of tweets (e.g., mentioning other users) and the emotional profiles of the feed. Concerning this latter aspect, the occurrences of emotion words in aggregated tweets can help to detect the tendency to be a fake news spreader. Thanks to the NRC Affect Intensity Lexicon [6], a manually annotated dataset of 6,000 English words collected with a technique called best–worst scaling (BWS), an intensity value for eight emotions can be derived.

- RTTR: root type-token ratio is a measure commonly used in NLP to assess the complexity of a text;
- mentions: number of mentioned users in the Twitter feed. Since the training set has been anonymized, it is impossible to have an idea of the variability and the type of mentioned users;
- replies: number of replies in the Twitter feed;
- urls: number of URLs in the Twitter feed;
- hashtags: number of hashtags in the Twitter feed;
- emoticon: number of emoticons in the Twitter feed;
- emphatic?: number of question marks in the Twitter feed;
- emphatic!: number of exclamation marks in the Twitter feed
- rich_people: the sum of occurrences of rich people's names in the Twitter feed. The list is composed by the world's highest-paid celebrities according to Forbes;

– all_emotion: the sum of values for all lemmas associated with all the emotions in the Twitter feed;
– fear: the sum of values for all lemmas associated with this emotion;
– trust: the sum of values for all lemmas associated with this emotion;
– anger: the sum of values for all lemmas associated with this emotion;
– sadness: the sum of values for all lemmas associated with this emotion;
– joy: the sum of values for all lemmas associated with this emotion;
– disgust: the sum of values for all lemmas associated with this emotion;
– anticipation: the sum of values for all lemmas associated with this emotion;
– surprise: the sum of values for all lemmas associated with this emotion.

To understand which features could be more discriminative for the two classes, a correlation analysis between each feature and the class value is proposed in Table 1.

features	**FNS_En**	**FNS_Es**
#mentions	-0.20	-0.28
#hashtags	-0.08	-0.30
#urls	0.04	0.26
#replies	-0.16	-0.25
#emoticons	-0.09	-0.13
#emphatic?	-0.07	-0.29
#emphatic!	-0.13	0.03
#RTTR	0.30	-0.22
#rich_people	0.24	0.14
#emotions_all	0.18	-0.05
#fear	0.24	-0.04
#trust	0.06	-0.06
#anger	0.26	-0.03
#sadness	0.20	0.03
#joy	-0.12	0.12
#disgust	0.19	0.004
#anticipation	-0.11	0.001
#surprise	0.22	0.08

Table 1. Pearson correlations between each feature and training set classes.

3.2 Experiments on the training set

In Table 2, the accuracy for different combinations of features on the training set is reported, applying random forest regressor and evaluating with 10-cross fold validation. Random forest regressors are not deterministic; for this reason, the mean accuracy for ten runs is reported.

– all features: 'rttr','mentions','urls','hashtags', 'replies','rich_people', 'emoticons', 'emphatic?', 'emphatic!', 'emotions_words', 'emotions_fear', 'emotions_trust', 'emotions_anger','emotions_sadness', 'emotions_joy', 'emotions_disgust','emotions_anticipation', 'emotions_surprise'

– communicative features: 'rttr','mentions','urls','hashtags', 'replies','rich_people'
– stylistic features: 'emoticons', 'emphatic?', 'emphatic!'
– emotions words: 'emotions_fear', 'emotions_trust', 'emotions_anger','emotions_sadness', 'emotions_joy', 'emotions_disgust','emotions_anticipation', 'emotions_surprise'
– best features: 'hashtags', 'emotions_fear', 'emotions_sadness'

features	FNS_En	FNS_Es
best features	0.702	0.698
emotion words	0.635	0.527
stylistic features	0.517	0.575
communicative features	0.629	0.687
all features	0.687	0.713

Table 2. Accuracy results for the training set.

3.3 Results on the test set

The test set of PAN @ CLEF 2020 competition *Profiling Fake News Spreaders on Twitter* is composed by 400 Twitter feeds. The best model, including number of hashtags and resulting from 10 cross-fold validation on the training set has been used for the regression-as-classification on the test set. Results on the test set are reported in Table 3.

dataset	Accuracy	Accuracy baseline
FNS_En	0.58	0.74
FNS_Es	0.5150	0.79

Table 3. Accuracy results for the test set.

4 Conclusions

In this paper, the methodology used to identify fake news spreaders for the PAN @ CLEF 2020 competition *Profiling Fake News Spreaders on Twitter* is described. After explaining data's pre-processing and the features extracted to classify fake news spreaders, a regression-as-classification approach is proposed; it represents being a fake news spreader as a gradable property. Performance (accuracy) on the training and the test set with the different sets of features is reported. The accuracy is below the baseline provided for this task and not in line with the results obtained on the trained set with the same methodology. As a consequence, further investigations are needed.

References

1. Alexander, D.E.: Social media in disaster risk reduction and crisis management. Science and Engineering Ethics, 20 pp. 717–733 (2014). https://doi.org/10.1007/s11948-013-9502-z
2. Breiman, L.: Random forests. Machine Learning, 45 pp. 5–32 (2001)
3. Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 877–880. SIGIR'19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3331184.3331285, https://doi.org/10.1145/3331184.3331285
4. Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., Derczynski, L.: SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 845–854. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2147, https://www.aclweb.org/anthology/S19-2147
5. McCreadie, R., Richard, M.C., Iadh, O.: Crowdsourced rumour identification during emergencies. In: Proceedings of the 24th International Conference on World Wide Web. pp. 965–970 (2015)
6. Mohammad, S.: Word affect intensities. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://www.aclweb.org/anthology/L18-1027
7. 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
8. Scheufele, D.A., Krause, N.M.: Science audiences, misinformation, and fake news. PNAS **16**, 7662–7669 (2019). https://doi.org/10.1073/pnas.1805871115
9. 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 (2018)
10. Vo, N., Lee, K.: Learning from fact-checkers: Analysis and generation of fact-checking language. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 335–344. SIGIR'19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3331184.3331248, https://doi.org/10.1145/3331184.3331248
11. Wardle, C.: Understanding Information Disorder. https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf?x76701 (2019)