# Profiling Fake News Spreaders: Personality and Visual Information Matter

Riccardo Cervero[1], Paolo Rosso[2], and Gabriella Pasi[1]

[1] Università degli Studi di Milano-Bicocca
[2] Universitat Politècnica de València
{r.cervero@campus., gabriella.pasi@}unimib.it, prosso@dsic.upv.es

**Abstract.** Fake news are spread by exploiting specific linguistic patterns aimed at triggering negative emotions and persuading the consumers. A way to contrast this phenomenon is to analyse the psychological factors underlying consumers' vulnerabilities. This paper is situated in this research context: first, we study the correlation between psycholinguistic patterns in user's posts and the tendency to spread false information. Moreover, since online contents exploit multimedia information, a methodology aimed at profiling the authors based on the images they share is employed. The reported experiments show that the proposed method, which considers both text-related and image-related features, outperforms the results of state-of-the-art approaches.

**Keywords:** Author Profiling · Personality Traits · Visual Information.

## 1 Introduction

Social Media platforms on the World Wide Web allow to easily provide a wide range of users with potential information. However, it is evident how an irresponsible use of these open systems may cause damages to the virtual community itself. This is the case of so-called fake news, an increasingly debated phenomenon described as false articles intentionally fabricated to mislead the audience [1]; they are able, for instance, to polarize public opinion, or to deceive non-expert readers about scientific issues. The growing use of social networks as a primary source of information has created non-intermediated contexts where the evaluation of credibility is left to users' judgment, which is however compromised by the difficulty to deal with unfamiliar topics. Moreover, the Social Web also favors strong peer-to-peer connections, fostering closed and toxic virtual environments like "echo chambers", whose main characteristic is the correlation between the intensity of user engagement and the degree of negative emotional polarity [2]. Shu et al. [3] highlighted how fake news exploit consumers' vulnerabilities, triggering negative emotions and irrational reactions. Hence, effective tools turn out to be algorithms able to learn the biases that penalise human judgement and to generate content that exploits them. Thus, there is an impelling need to contrast online disinformation; one possible means is to detect the users who are

potential generators or sharers of fake news, by identifying the individual vulnerabilities at the basis of a lower capability to discern genuine content from fake one. Assuming that these vulnerabilities derive from psychological inclinations, this work aims to demonstrate that "fake news spreaders" are associated with specific personality traits. Therefore, after extracting personality characteristics from users' texts, we both evaluate their impact on the tendency to spread false content and test their effectiveness for the task of binary classification of users into real or fake news spreaders. However, as content flows quickly in microblogs, users' attention may be initially attracted by the visual elements of the posts. It is possible that images embedded in fake news attempt to exploit cognitive vulnerabilities, and thus they may present specific patterns. For this reason, we also report in this paper the outcomes of investigating the impact of visual features on fake news spreaders' profiling. In conclusion, the main contributions of this paper are the following. Firstly, inspired by Giachanou et al. [4], we evaluate the effectiveness of psycho-linguistic features to perform a classification of users into real and fake news spreaders. As a second task, inspired by [5], we also analyze the effectiveness of visual features - alone or mixed with personality information - to classify the authors. Lastly, we verify the feasibility of improving the effectiveness of state-of-the-art approaches for fake news detection by incorporating and/or replacing the proposed personality and visual information into the best models at the Author Profiling Task at PAN 2020[3].

The rest of the paper is organised as follows: Section 2 presents related works about the author profiling perspective; Section 3 introduces the PAN 2020 dataset and the two best performing solutions; Section 4 illustrates the methods of psycho-linguistic features extraction; Section 5 explains how visual information is obtained; Section 6 describes all the experiments carried out to evaluate the effectiveness of the aforementioned research contributions, whose obtained results are commented in the last Section 7.

## 2 Related Work

Fake news spreaders detection is an increasingly investigated research topic, which is more commonly tackled by means of data-driven approaches. Popular solutions are based on stylometric analysis - which aims to identify which style fits the category of "fake news spreaders", as in [6] -, or the extraction of lexicon-based emotional dimensions - the same on which Giachanou et al. [7] train an LSTM model. In particular, the employed features at the Author Profiling task at PAN 2020 can be divided into four categories [8]: *(i)* words or characters n-grams, *(ii)* stylistics, *(iii)* embeddings, *(iv)* personality and emotions, or combinations thereof. The best solutions respectively exploited a combination of n-grams and stylistic features (Buda & Bolonyai [9]) and only n-grams (Pizarro [10]). Regarding the personality descriptors, the reference point of this work is what has been done by Giachanou et al. in [4], covered in the Section 4. Previous alternatives were the Myers-Briggs Type Indicator [11], or the manual

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

compilation of questionnaires. Images are less frequently considered for the author profiling task, and the combination of visual and personality information is still under-explored. The reference point, in this case, is the approach aimed at extracting the visual features in [5] (Section 5).

## 3 PAN 2020: Profiling Fake News Spreaders on Twitter

The 2020 edition of the PAN event came with a shared task [8] aimed to inquire the feasibility of detecting authors who shared fake news in their past timeline in a bilingual perspective, i.e. considering both English and Spanish tweets. Two datasets, provided separately for each language, were generated as explained below. After selecting news labelled as fake on debunking websites, the organizers downloaded and manually labelled the tweets related to them as content supporting the false information, or vice versa. Thus, users in the sample who had shared at least one tweet supporting a fake news were labelled as "fake news spreader", and only the ones with the highest count were included in the final dataset, together with the same number of randomly selected "real news spreaders". In the end, the datasets are generated by collecting the last 100 tweets from each user's timeline, discarding those directly related to the fake news considered above, so as to avoid biases. On these datasets, the best average performance has been achieved, with equal merit, by Buda & Bolonyai [9] and by Pizarro [10]. In details, Buda-Bolonyai's model provided the highest accuracy on the English dataset (0.75), while Pizarro obtained the best result on Spanish tweets (0.82). Buda & Bolonyai's solution [9] is structured as follows. Firstly, four baseline classifiers (Logistic Regression, Support Vector Machine, Random Forest, and the gradient boosting algorithm XGBoost) undergo a training process consisting in an extensive grid search of the optimal combination among text pre-processing methods, vectorization techniques and baseline parameters. In details, the authors experimented different ranges for words n-grams. Then, another XGBoost algorithm is trained on user-wise statistical indicators: *(i)* minimum, maximum, mean, standard deviation and range of the length - both in words and in characters - of the tweets; *(ii)* number of retweets and mentions by the author; *(iii)* count of additional elements: URLs, hashtags, emojis and ellipses; *(iv)* lexical diversity calculated as the type-token ratio of lemmas. Buda & Bolonyai have preferred cross-validation techniques to prevent overfitting while optimizing the parameters of the baselines, instead of a single hold-out. Lastly, the five submodels are trained to determine the probability of being a fake news spreader, and then they are stacked together through the best ensemble method chosen among *(i)* Majority Voting, *(ii)* Linear Regression, and *(iii)* Logistic Regression, which turned out to be the most reliable. The training of the ensemble model has been performed on the approximation of the predictions distribution, obtained by refitting the sub-models on different chunks of the training set. Pizarro [10] performed an optimization of the parameters of a Linear Support Vector Classifier trained on combinations of word and character n-grams, and experimenting with twelve pre-processing pipelines, based on mixtures of four basic operations: *(i)*

downcase all the letters; *(ii)* replace numbers, URLs, users' name and hashtags with tokens; *(iii)* replace emojis with word representation; *(iv)* reduce number of repeated characters. The final linguistic features derive from the calculation of the Term Frequency - Inverse Document Frequency (TF-IDF) weight for each n-gram.

## 4 Personality Information

Giachanou et al. [4] originally proposed a method to classify users into "fake news spreaders" and "fact checkers", i.e. those interested to share posts that refute false information with evidences. In this work, instead, we aim at a classification into fake news "speaders" and "non spreaders". Their CheckerOrSpreader architecture is composed of a Convolutional Neural Network built upon two components: *(1)* one aimed at defining word-embeddings vectors from a pre-trained GloVe model, and *(2)* one aimed at eliciting the psycho-linguistic information that can describe users' personality. To obtain the latter, two approaches have been used simultaneously: *(i)* use of the LIWC software [12], mapping the text into 73 "psychologically-meaningful categories"; *(ii)* the Five-Factor Model (FFM) [13], which quantifies the evidence of a particular trait or disorder in user's text, considering five basic factor: openness to experience, conscientiousness, agreeableness, extraversion, and neuroticism. A personality score is then derived with Neuman and Cohen's method [14], computing the semantic similarity between the context-free embedding representations of both input text and a set of benchmark adjectives empirically observed as to be able to encode the essence of personality. The aforementioned components are then combined with further sets of features: *(1)* eight emotional dimensions and two related to the sentiment polarization (both through the NRC lexicon)[4]; *(2)* Bag-Of-Words vectors. The CheckerOrSpreader model was trained and tested on the two PAN datasets, obtaining an accuracy repectively equal to 0.52 and 0.51 for English and Spanish datasets. This result will be useful for subsequent comparisons.

## 5 Visual Information

In [5], the authors mined features from the images embedded in the tweets, by using pre-trained neural networks. The work presented in this paper follows a similar extraction methodology, although from an author profiling perspective. This new approach offers an average description of all the images posted by each user in the sample, and subsequently it evaluates to which of the two target classes this description can correspond. In details, the set of non-duplicated images scraped from each user's texts are passed to five models, pre-trained on the popular ImageNet dataset - VGG16, VGG19, ResNet50, InceptionV3, Xception. After applying an average pooling operation, five vectors per image are obtained. Finally, the five compressed representations - one per neural network -,

[4] https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

from which it is possible to draw the typical characteristics of the visual contents posted by a user, are obtained by averaging all the vectors per image. In case no images were available for a particular author, vectors of zero have been assigned.

## 6 Experiments and Results

All the experiments have been performed on the same training and test sets provided at the Author Profiling Task at PAN 2020. All the results, organized with reference to the considered research issues, are available at the following link: `github.com/results`. The evaluation metric considered is the Accuracy.

First of all, to verify the effectiveness of psycho-linguistic information, we tested all the possible combinations between the LIWC features and the Five Factor model features - separately or jointly -, mixing with the emotional dimensions and the BOW vectors mentioned in Section 4, as well as the statistical features implemented by Buda-Bolonyai (Section 3). The predictive architectures tested are: Logistic Regression (LogReg), Convolutional Neural Network (CNN) - like the CheckerOrSpreader model requires - and a Long Short-Term Memory (LSTM). The last two also consider as input the vector representation of the tweets provided respectively by the pre-trained GloVe model and a pre-trained FastText model for the English and Spanish datasets, both fine-tuned to better capture the semantic contexts. Looking at the results displayed in the Tables A (available `here`) and B (available `here`), the best solution for both languages appears to be a Logistic Regression trained on the mix of personality scores and TF-IDF values, offering an accuracy of 0.69 for English and 0.75 for Spanish. Despite the fact that these values are respectively lower w.r.t. Buda-Bolonyai's performance (0.75 on English users) and Pizarro's result (0.82 on the Spanish dataset), the difference w.r.t. Buda-Bolonyai's accuracy is not statistically significant with a confidence level set at 95%. This confirms that personality scores derived by FFM - in combination with a BOW approach - are powerful enough to significantly conform the state-of-the-art performances on the author profiling task in case of English text. Personality scores without BOW vectors always offer poorer results, but, in the English case, still better than the accuracies produced by the LIWC features alone. In contrast, this latter software-generated representation outperforms the FFM on Spanish text. It is then also possible to conclude that emotional and Buda-Bolonyai's features, in combination with personality information, make a little contribution to the accuracy result. Finally, it is important to note that deeper models like CNN or LSTM always give worse results than Logistic Regression, probably because this latter is able to intrinsically manage the strong collinearity among variables in a better way than the two others architectures, and in general its performance is not penalised by a small amount of data, as is the case with neural networks.

The evaluation of the usefulness of a user's "visual profile" for the given task - second goal of the project - consisted in the experimentation of all possible

combinations among the five vector representations from each truncated neural network, and, at a later time, mixing also with the psycho-linguistic components, emotional dimensions and Bag-Of-Words. The tests have been carried out by training a Logistic Regression, since, as aforementioned, this ensures efficient management of the strong multicollinearity among the visual features. Observing the results in Tables C (available `here`) and D (available `here`), in both cases the best solution remains the union of the linguistic patterns extracted from the LIWC software with the visual information, even if the results on the two datasets (0.675 for English and 0.706 for Spanish) are lower than the best performances reported for the PAN task in 2020. It is important to note that in the Spanish case the VGG16 vector representation appears to be the only useful one. Finally, although it may seem that visual information alone offers poor accuracy (0.59 with a VGG16-Xnception combination on English users and 0.553 with VGG16 vector for Spanish ones), it is necessary to consider that these solutions, actually, still manage to exceed the result achieved by the original CheckerOrSpreader model on the same PAN test sets (respectively 0.52 and 0.51 for English and Spanish datasets), even without considering any textual information at all. In the first case, this difference is even statistically significant with a 95% confidence level.

Regarding improvements to state-of-the-art models, it is worth mentioning that, to reduce computational weight and training time, only the best performing combinations between visual and psycho-emotional information have been tested. As far as variations on Buda-Bolonyai's model, we maintained the simultaneous training of the four baselines on word n-grams. The variations, instead, concerned the features set the XGBoost algorithm is trained on, and the trial of both Logistic Regression and Linear Regression as an ensemble method. Focusing on English dataset, we can see that any replacement and integration of visual/personality information in the ensemble model improves the original result (as visible in Table E, available `here`). The maximum accuracy (0.775) is reached with the combination of the baselines trained on N-grams plus an XGBoost model fed with personality scores and VGG16-Xnception vectors. The only exception - an accuracy worse than the original one - is found when only integrating the Five Factor Model representation. In the Spanish sample, the opposite is observed: any modification worsens the original result (as seen in Table F, available `here`). With regard to Pizzaro's system, since it was iteratively trained only on mixtures of n-grams extracted from pre-processed text, the modifications consisted in simple concatenations of the new features sets - including visual, statistical and psycho-emotional features - to the TF-IDF weights originally considered. However, it has been necessary to estimate the personality scores only once with the original text preparation performed by Giachanou et al. [4]. Since the FFM paradigm compresses input text to compute the similarity with the vectors of the benchmark adjectives, variations in text preparation – searching for an optimal pipeline, as Pizarro's original system does - could penalize the result. From Tables G (available `here`) and H (available `here`), it appears that, for both datasets, integrations almost always lead

to a performance worsening. The exception on the English dataset occurs with the only addition of personality scores (with an increases from 0.735 to 0.76). On the Spanish dataset, the only improvement in the accuracy is due to the concatenation of the VGG16 vector: the result rises from 0.82 to 0.832.

## 7 Conclusions

Fake news is an increasingly debated phenomenon due to the dramatic influences it has on both virtual and real communities. It is, thus, of primary importance that scientific research is concerned with countering the spread of false information. In this paper, we test the effectiveness of personality information and visual features for profiling fake news spreaders on Twitter. To summarise the results obtained from the performed experiments, Tables 1 and 2 show the accuracy measures achieved by the respective best combinations of features sets and predictive models, on both English and Spanish corpora. From these Tables, it appears that the fake news spreader detection task can be addressed more effectively with a combination of N-grams, personality information and visual features, in both datasets. Therefore, the obtained results demonstrate the relevance of the visual and personality information proposed. In both languages, the second best solution combines visual information with textual features. A second consideration can thus be made on the effectiveness of visual features: although they offer worse results if used alone, when combined with N-grams they always obtain a better performance w.r.t. the mix of text and personality information. In general, even in combination with psycho-linguistic features, visual information offers good results. Then, personality scores, modeled together with BOW by a Logistic Regression, offer a result statistically not inferior to state-of-the-art solutions for English only. In particular, we observe that the most powerful psycho-linguistic features in both languages are offered by the Five Factor Model. However, when combined with the LIWC patterns, it often penalizes the result. Moreover, in this context it was noted that it is advisable to use less complex models like Logistic Regression. Finally, it is possible to conclude that the integration of visual/personality information allows to improve the performance of state-of-the-art models in many cases.

Table 1: Best overall solution on the English dataset.

| Combination | Model | Features | Accuracy |
|---|---|---|---|
| TXT+PERS+IMG | LinReg Ensemble | N-grams + FFM + VGG16, XNC | **0.775** |
| TXT+STAT | *Buda-Bolonyai's* | N-grams + Stat. | 0.75 |
| TXT | *Pizarro's* | N-grams | 0.735 |

Table 2: Best overall solution on the Spanish dataset.

| Combination | Model | Features | Accuracy |
|---|---|---|---|
| TXT+PERS+IMG | Linear SVC | N-grams + FFM + VGG16 | **0.832** |
| TXT | *Pizarro's* | N-grams | 0.82 |
| TXT + STAT | *Buda-Bolonyai's* | N-grams + Stat. | 0.805 |

## 8 Acknowledgements

This work is funded by Project 2020-ATE-0632, "Definition of models and systems for the representation, management and analysis of information and knowledge", University of Milano Bicocca, and supported by the MISMIS-FAKEnHATE research project on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

## References

1. Allcott H., Gentzkow M.: Social media and fake news in the 2016 election. In: National Bureau of Economic Research (2017)
2. Del Vicario M., Vivaldo G., Bessi A., Zollo F., Scala A., Caldarelli G., Quattrociocchi W.: Echo chambers: Emotional Contagion and Group Polarization on Facebook. In: Scientific Reports 6 (2016)
3. Shu K., Sliva A., Wang S., Tang J., Liu H.: Fake News Detection on Social Media: a Data Mining Perspective. In: ACM SIGKDD Explorations Newsletter 19 (2017)
4. Giachanou A., Ríssola E., Ghanem B., Crestani F., Rosso P.: The Role of Personality and Linguistic Patterns in Discriminating Between Fake News Spreaders and Fact Checkers. In: Proc. 25th Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2020, LNCS(12089), 181-192 (2020)
5. Giachanou A., Zhang G., Rosso P.: Multimodal Fake News Detection with Textual, Visual and Semantic Information. In: Proc. 23rd Int. Conf. on Text, Speech and Dialogue, TSD-2020, Springer-Verlag, LNAI(12284), 30-38 (2020)
6. Afroz S., Brennan M., Greenstadt R.: Detecting hoaxes, frauds, and deception in writng style online. In: ISSP'12 (2012)
7. Giachanou A., Rosso P., Crestani F.: Leveraging Emotional Signals for Credibility Detection. In: Proc. of the 42nd Int. ACM SIGIR Conf. on Research and Development in IR, 877–880 (2019)
8. 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-WS.org, vol. 2696 (2020)
9. Buda J., Bolonyai F.: An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (2020)
10. Pizarro J.: Using N-grams to detect Fake News Spreaders on Twitter. In: Cappellato L., Eickhoff C., Ferro N., Névéol A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (2020)
11. Briggs-Myers, I., Myers, P.B.: Gifts differing: Understanding Personality Type. Davies-Black Publishing. (1995)
12. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The Development and Psychometric Properties of LIWC 2015. Tech. rep. (2015)
13. John, O.P., Srivastava, S.: The Big-five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In: Handbook of Personality: Theory and Research, 102–138 (1999)
14. Neuman Y., Cohen Y.: A Vectorial Semantics Approach to Personality Assessment. In: Scientific Reports 4(1) (2014)