Profiling and Detecting Fake News Spreaders on Social Media

Modelado e Identificación de Perfiles de Usuarios Difusores de Noticias Falsas en Redes Sociales

María S. Espinosa
Natural Language Processing and Information Retrieval Group
Universidad Nacional de Educación a Distancia
mespinosa@lsi.uned.es

Abstract: Nowadays our society is subjected to a massive spread of digital information. This spread has led to a great amount of false information, such as rumours, *fake news*, and extremely biased news, being shared and consumed by Internet users every day. Due to the great impact of *fake news* on politics, economy and health, it is becoming essential to design tools for the automatic verification of the veracity of online information. From this point of view, this proposal aims to research about the detection and classification of information that will help identifying internet users as *fake news* spreaders or reliable sources.
Keywords: fake news, information verification, user profiling

Resumen: Hoy en día la sociedad está expuesta a una difusión masiva de información digital. Esta difusión ha provocado que los usuarios de Internet compartan y consuman una gran cantidad de información falsa como son los rumores, las noticias falsas o la información subjetiva y extremadamente sesgada. Debido al gran impacto que tienen las noticias falsas en la política, la economía y la sanidad, la creación de herramientas para la verificación automática de la información online resulta esencial. Desde esta perspectiva, esta propuesta tiene el objetivo de investigar un sistema de detección y clasificación de información que ayude en la identificación de usuarios como difusores de noticias falsas o fuentes de información fiable.
Palabras clave: noticias falsas, verificación de información, perfilado de usuarios

1 *Motivation*

Due to the increasing amount time spent in social networks in the past years, people have changed their ways of news consumption, moving from traditional media such as TV and newspapers, to social networks, such as Facebook or Twitter.

According to the United States *Pew Research Center*, in 2016[1] approximately 62 % of adult Americans reported to get their news via social media, whereas in 2012 this value was 49 %, going up to 68 % in 2018[2].This rapid increase is associated with elements such as a lower price and a greater speed and immediacy in the dissemination of news. Thus, social networks have become a fundamental publication tool for journalists (Tolmie et al., 2017).

Despite these advantages, the massive spread of digital information to which our society is subjected nowadays has led to a great amount of false or extremely biased information being shared and consumed by Internet users every day.

One of the most notorious examples of the influence that *fake news* have on society is the 2016 U.S. Presidential Elections. Online political discussion was strongly influenced by social media users and bots spreading misinformation, which potentially altered public opinion and endangered the integrity of the elections. Furthermore, a study conducted in 2017 revealed that, in the 3 months prior to the elections, 115 pieces of pro-Trump *fake news* were spread, and they reached 30 million shares, while 41 pieces of pro-Clinton

[1] https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

[2] https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

Proceedings of the Doctoral Symposium on Natural Language Processing from the PLN.net network (PLNnet-DS-2020)

CEUR Workshop Proceedings (CEUR-WS.org)

fake news were spread with 7,6 million shares (Allcott and Gentzkow, 2017).

Fake news can have a great impact on economy as well. In 2017, news claiming that Barack Obama was injured in an explosion wiped out $130 billion in stock value (Rapoza, 2017).

Sometimes, the spread of false information has the objective of increasing the panic and creating chaos, as is the case of the "Pizzagate", which led a man to assault a restaurant with a rifle in 2016. *Fake news* stating that the restaurant was harboring young children as sex slaves as part of a child-abuse ring led by Hillary Clinton were widespread across the web (Kang and Goldman, 2016).

These are examples of the high cost associated to the spread of *fake news*: the absence of control and verification of the information, which makes social media a fertile ground for the spread of unverified or false information.

With this in mind, we can affirm that the magnitude, diversity and substantial dangers of *fake news* and, in more general terms, the disinformation circulating on social media is becoming a reason of concern due to the potential social cost it may have in the near future (Allcott and Gentzkow, 2017).

For all these reasons, the task of fake news detection in general and, in particular, detecting fake news spreaders, have become a cornerstone problem nowadays. Thus, the main objective of this proposal focuses on analysing the role of user profiles features for such a task.

2 Background and Related Work

The use of the expression *fake news* for identifying false information spread online has become very popular, especially after the 2016 U.S. Presidential Campaign. The most extended definition applies to news pieces are intentionally written to mislead or misinform readers, but can be verified as false by means of other sources (Conroy, Rubin, and Chen, 2015).

The intention behind the creation and dissemination of *fake news* often has a political or economic component. Given the crucial role that the spread of *fake news* plays in our current society, research on this topic is developing significantly. In fact, the number of published papers indexed in the the Scopus database concerning the topic of *fake news* has increased considerably from less than 20 in 2006 to more than 200 in 2018 (Zhou and Zafarani, 2018). These works concentrate on understanding how false information spreads through social media, and how can it be efficiently detected in order to reduce its negative impact on society. This task has been approached from different perspectives, such as *Natural Language Processing* (NLP), *Data Mining* (DM), and *Social Media Analysis* (SMA).

In many cases the task is treated as a binary classification problem where a news piece is classified as fake or real. However, there are cases in which this classification may not be adequate since the news could be partially true and partially false. For this reason, systems capable of multi-class classification have also been proposed (Rashkin et al., 2017).

In the field of Natural Language Processing, research has been focusing on the detection and intervention of *fake news* using techniques such as *Machine Learning* and *Deep Learning* (Ruchansky, Seo, and Liu, 2017), and taking into account:

- Content-based features contain information that can be extracted from the text, such as linguistic features.
- Context-based features contain surrounding information such as user characteristics, social network propagation features, or users' reactions to the information.

Detecting *fake news* in the context of social media presents characteristics and challenges that result in content-based methods not being effective on their own. *Fake news* are intentionally created to deceive, making it difficult identify them only from their textual content. For this reason, it is common to use surrounding information such as the way in which they are disseminated and the behavior of the users involved in this dissemination, as well as information related to the author of the news (Shu et al., 2017).

Recent research has demonstrated that studying the correlation of the user profile and the spread of *fake news* works for the identification of those users mere likely to believe *fake news* and for the differentiation of those more likely to believe real news(Shu, Wang, and Liu, 2018). Approaches considering context-based features in combination with content-based features have been gaining popularity in the past years, due to the promising results obtained in recent studies (Shu et al., 2019). Mainly three aspects of this type of information can be studied:

- User information, such as location, age, number of followers, etc.
- The responses generated by *fake news*, which can stand as an important source of detection not only because users use responses to express their opinions but also because they can help in the construction of a credibility index for users (Jin et al., 2016).
- The social networks through which the news disseminate. The study of the networks through which the information is propagated has special relevance since the rapid diffusion of these networks is used to reach the maximum number of users in the shortest possible time.

3 Research Proposal and Main Hypotheses

This research proposal is centered in analysing the role of user profiles facing the task of mitigating the spreading of *fake news* focusing on the identification of user profiles prone to *fake news* spreading. It is based on the following hypotheses:

1. **There are users prone to the spread of *fake news* in social networks.** The majority of studies in this area are focused on the detection of *fake news* based on the content of the news itself. This research project aims to demonstrate that putting the user in the center of the process will help in the detection of *fake news*. As a consequence of this, we will be able to identify those users prone to the creation and spreading of *fake news*.
2. **There is a set of features that differentiates *fake news* spreaders.** This hypothesis relies on the first hypothesis. Given a set of users prone to the spread of *fake news*, we want to identify the set of features that characterize them in order to be able to identify and differentiate *fake news* spreaders from real news spreaders.
3. **Establishing the difference between user-created and user-shared content will reveal more accurate features of the user's online behavior.** The main contents that users share in social media can be divided in: (1) content created by the user, and (2) content created by others. This hypothesis states that the individual analysis of these groups of contents will reveal more precise features of the user profiles.

For the evaluation of these hypotheses, this proposal will perform the following contributions:

- It will produce a study of the relation between user profiles and *fake news*, which will lay the grounds for its use in the detection of *fake news* spreaders.
- It will produce a comparative analysis of the different features identifying *fake news* spreaders from different dimensions.
- It will demonstrate the utility of moving the user to the centre of the process of mitigation of the *fake news* spreading.

4 Objectives

The main objective of this PhD research project is to mitigate the spread of *fake news* online through the study and identification of user profiles based on features of the own textual content and user profiles features. Both type of features will be tackled from two different dimensions: user-created and user-shared content. In order to achieve this general objective, we have set the following specific goals:

1. **Problem formalization.** As a preliminary step, We need to formalize the problem of detecting *fake news* spreaders on social networks based on content and user profiles features.
2. **Data collection.** We will create a dataset of real data from social networks where false and real news have been spread.
3. ***Fake news* spreader identification.** From the data collected in objective 2, an identification of the users who share *fake news* will be conducted in order to build a representative set based on their user profiles.

4. **User profile analysis.** The analysis will be conducted according to hypothesis 3 and from the two identified dimensions: user-created content and user-shared content, separately.

5. **Explore learning approaches.** Explore and define different approaches based on machine learning and deep learning.

6. **Empirical evaluation.** We will measure the importance and usefulness of each of the identified features, analyzing their contribution for the classification of user profiles.

7. **Result analysis.** Finally, we will carry out an analysis and dissemination of the results obtained in terms of generated resources and evaluation data, as well as in the form of scientific publications and participation in evaluation tasks within the project's area.

5 *Current Work*

The work here proposed is in its initial stages of development. As part of the development of this proposal we presented a preliminary model developed for the detection of *fake news* spreaders on Twitter. The model was evaluated at the *Profiling Fake News Spreaders on Twitter* task on Author Profiling at the PAN@CLEF 2020 competition.

5.1 Feature Engineering

Our model approached the problem of identifying *fake news* spreaders from some of the dimensions discussed in this proposal. Due to the dataset restrictions, some aspects of the user profiles could not be evaluated because the training data for the task was obfuscated for privacy reasons. However, we managed to build up a model that considered personality traits, linguistic features, social media activity data, and shared media features.

The psychological features were extracted using a third-party API developed by Symanto[3]. The documents containing the aggregated tweets for each user were sent to the API in order to retrieve the values of their (1) personality traits, (2) their communication styles, and (3) the sentiment analysis of their text.

For the extraction of linguistic features, a natural language pipeline called Polyglot was used (Al-Rfou, Perozzi, and Skiena, 2013). This library is built using distributed word representations (word embeddings) in conjunction with traditional NLP features in order to solve NLP tasks. For our model's set of features we choose 12 POS tagging metrics, 3 named entity recognition metrics and total word count.

The analysis of the activities of the users in *Twitter* was restricted by the data obfuscation. Therefore, only 4 metrics were recorded from the actions of the user within *Twitter*: the number of mentions, the URL number, the number of retweets and the number of hashtags. The values of these metrics were counted from the total aggregation of tweets of each user.

Based on the assumption of hypothesis 3 being true, we separated tweets from retweets and applied the last set of measurements only to the retweet subset. This category consisted on headline analysis and we studied if there are specific message characteristics that accompany fake news articles being produced and widely shared. Recent studies suggest that not only these characteristics exist, but also that some of them can be found in the headline of the news article (Horne and Adali, 2017). Therefore, we took the 3 most significative characteristics differentiating fake news headlines from real news headline and applied them to the text in our dataset.

5.2 Results

For the creation of our model we first did some experiments in order to select the most important features as well as the best performing algorithms[4] We tested the model taking into account the set of features available in each category separately, and we also tested the possible combinations of the features to evaluate their performance on the data. After comparing the results obtained with the different categories and classifiers, we trained the model using a combination of all categories. With regards to the classification algorithm, the *Random Forest Classifier* was chosen as the algorithm to train our model due to the good results obtained in all the experiments.

There were two evaluations for our model. On the one hand, there was an early-bird submission evaluation for the task and, on the other hand, there was the final sub-

[3] https://symanto-research.github.io/symanto-docs/

[4] All the experiments and results can be found in a notebook uploaded to this github repository.

mission evaluation. We participated in both evaluations, first with and early model and then with a final model. The evaluation results can be found in table 1.

Data	Model	Phase	Accuracy
test	early	experim.	0.67
test	final	experim.	0.68
eval.	early	*early-bird*	0.67
eval.	final	final	0.64

Tabla 1: Early and final model evaluation results throughout the different evaluation phases: experimentation, *early-bird* submission, and final submission.

With regards to the results in English language our team was positioned 45th form 66 participants. Furthermore, if we aggregate the results, that is, if we count all the participants with the same results as just one participant, our result would be 16th from 33 participants.

6 *Methodology and Experiments*

This PhD research project aims to create a model for user profiling, which encompasses many different aspects of the user profile. For this reason, in order to define our methodology, a set of experimental stages have been defined.

6.1 Data collection and formalization of the problem

The first set of experiments aims to analyze and identify the main features of the proposed problem in order to study their strengths and weaknesses. Moreover, published datasets in the field of *fake news* will be collected and analyzed in order to apply them to the problem stated in this research project. Specifically, we will study the following datasets: FakeNewsNet (Shu et al., 2018), LIAR (Wang, 2017), CREDBANK(Mitra and Gilbert, 2015), and BuzzFace (Santia and Williams, 2018).

Once the data has been collected, the user profiles who have shared *fake news* will be distinguished of those who have not. The result will constitute the representative sets of *fake news* spreaders and real news spreaders, along with the set of collected features.

6.2 Identification and analysis of *fake news* spreaders profile traits

In this stage, user-created content will be differentiated from user-shared content in order to proceed with the analysis of the features that characterize *fake news* spreaders and real news spreaders. This analysis will be conducted based on two different dimensions: the social and demographic dimension and the linguistic dimension.

From a social and demographic point of view, we will focus on aspects studied in other related and interdisciplinary areas, such as psychology. Specifically, we will focus on extracting aspects from their online content, such as personality, emotion, age, gender, and location. We will also study their activity in social media, analyzing aspects such as the number and type of connections they have with other users, the detection of communities within the network, the number of messages published, number of followers, etc.

Regarding the linguistic dimension, we will explore and evaluate user traits extracted from linguistic indexes and speech representation of texts. For this, aspects of theories based on discourse analysis and text coherence will be studied, such as the Rhetorical Structure Theory (Mann and Thompson, 1988). Likewise, embedding-based text representation methods will be explored, such as Twitter2Vec (Vosoughi, Vijayaraghavan, and Roy, 2016) or Word2Vec (Mikolov et al., 2013).

6.3 Analysis and proposal of techniques based on machine learning and deep learning

After the definition of the user profile distinctive traits, a thorough study of the state-of-the-art *Machine Learning* classification algorithms will be conducted in order to learn to classify and identify user profiles as *fake news* or real news spreaders. Specifically, we will use the most popular classifiers, such as Random Forest, Support Vector Machine, Decision Trees, and Logistic Regression. As a result, we will obtain what type of classifiers work best when solving the task.

Next, we will focus on designing a new proposal for a learning mechanism based on *Deep Learning* techniques, taking into account the results obtained by the different algorithms experimented, as well as the state-of-the-art techniques in the field of *fake news* detection. We will concentrate in popular deep learning models that have been successfully used for implementing language models, such as Long Short Term Memory Recurrent Neural Networks(LSTM, RNN), and Convolutional Neural Networks (CNN)(Jang, Kim, and Kim, 2019).

7 Research Issues for Discussion

The present research proposal has the aim of examining how user profiles can be modeled in order to classify users as *fake news* spreaders or real news spreaders. In order to achieve this objective, the main task is to combine content and context features of the user to obtain the user profile model. The outlined approach raises the following issues for discussion:

- Which is the role of user profile information in the identification of fake news spreaders?
- Which is the most effective way to combine this information in order to achieve optimal results?
- Assuming we depend on existing tools and resources from other scientific areas, such as psychology and linguistics, how can we achieve a balance between the information they provide and the deficiencies that these tools have when combining several of them in a study?

Acknowledgements

This research project is funded by the European Social Fund through the Youth Employment Initiative (YEI 2019) under the grant PEJD-2019-PRE/TIC-15255 and the Spanish Ministry of Science, Innovation and Universities (DeepReading RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE).

Special thanks to Roberto Centeno, Álvaro Rodrigo and the NLP&IR UNED Research Group for their support.

References

[Al-Rfou, Perozzi, and Skiena2013] Al-Rfou, R., B. Perozzi, and S. Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

[Allcott and Gentzkow2017] Allcott, H. and M. Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

[Conroy, Rubin, and Chen2015] Conroy, N. K., V. L. Rubin, and Y. Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

[Horne and Adali2017] Horne, B. D. and S. Adali. 2017. 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*.

[Jang, Kim, and Kim2019] Jang, B., I. Kim, and J. W. Kim. 2019. Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, 14(8):e0220976.

[Jin et al.2016] Jin, Z., J. Cao, Y. Zhang, and J. Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Thirtieth AAAI conference on artificial intelligence*.

[Kang and Goldman2016] Kang, C. and A. Goldman. 2016. In washington pizzeria attack, fake news brought real guns. *The New York Times*.

[Mann and Thompson1988] Mann, W. C. and S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

[Mikolov et al.2013] Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Mitra and Gilbert2015] Mitra, T. and E. Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pages 258–267.

[Rapoza2017] Rapoza, K. 2017. Can 'fake news' impact the stock market? *by Forbes*.

[Rashkin et al.2017] Rashkin, H., E. Choi, J. Y. Jang, S. Volkova, and Y. Choi.

2017. 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*, pages 2931–2937.

[Ruchansky, Seo, and Liu2017] Ruchansky, N., S. Seo, and Y. Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806.

[Santia and Williams2018] Santia, G. C. and J. R. Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Twelfth International AAAI Conference on Web and Social Media*.

[Shu et al.2018] Shu, K., D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 8.

[Shu et al.2017] Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

[Shu, Wang, and Liu2018] Shu, K., S. Wang, and H. Liu. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435. IEEE.

[Shu et al.2019] Shu, K., X. Zhou, S. Wang, R. Zafarani, and H. Liu. 2019. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439.

[Tolmie et al.2017] Tolmie, P., R. Procter, D. W. Randall, M. Rouncefield, C. Burger, G. Wong Sak Hoi, A. Zubiaga, and M. Liakata. 2017. Supporting the use of user generated content in journalistic practice. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 3632–3644.

[Vosoughi, Vijayaraghavan, and Roy2016] Vosoughi, S., P. Vijayaraghavan, and D. Roy. 2016. Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1041–1044.

[Wang2017] Wang, W. Y. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

[Zhou and Zafarani2018] Zhou, X. and R. Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.