Multi-dimensional feature extraction for the identification of online harm spreaders

Daniela Godoy
ISISTAN (UNCPBA-CONICET)
Facultad de Ciencias Exactas, UNCPBA
Tandil, CP 7000, Argentina
daniela.godoy@isistan.unicen.edu.ar

***Abstract*—Identifying online harm spreaders, such as fake news or hate speech spreaders, has become an important problem to prevent the propagation of low-quality content online. In this work, several dimensions of the user behaviour are defined to analyze this type of users and a set of features are extracted within each dimension, leveraging mostly lexicons, in order to better characterize spreaders and help to its detection.**

***Index Terms*—Online harm spreaders, feature extraction, misinformation, hate speech.**

I. Introduction

Under the umbrella of online harm spreaders there is a variety of users acting maliciously to propagate harmful content on social networks. Although the most noticeable and more studied ones are fake news spreaders and hate speech spreaders, those users sharing rumors, conspiracy theories or different types of offensive comments, including ironic, abusive or discriminatory content, also fall into this category.

Several works in the literature have been centered on evaluating machine learning approaches to distinguish some specific type of harm spreader (e.g. fake news spreaders, hate spreaders, etc.) from regular users or users counteracting their actions [1]–[7]. In these works, the prediction of a user likelihood of being a spreader was assessed based on a varied set of features describing the user-generated content and behavior within the framework of a binary classification problem. In this context, features representing linguistic patterns, emotions, sentiment, among other aspects, have been used to classify some specific type of harm spreader acting online.

In this work, a large set of features spanning multiple dimensions to represent the behaviour and content generated by users is defined in order to study their prevalence in those users propagating some kind of harmful content. The hypothesis is that these dimensions can help to better characterize and, therefore, classify the different types of spreaders under the more general class of online harm spreaders. Instead of restricting the problem to a binary classification one focused on a reduced set of features, a more general representation of users is defined which can, therefore, serve several binary or multi-class classification tasks.

The rest of the paper is organized as follows. Section II discusses related works. Section III describes the dimensions and features used for characterizing spreaders. Data used for experimentation is described in Section IV and results are reported in Section V. Conclusions are discussed in Section VI.

II. Related Works

The problem of profiling spreaders has been mostly seen as a binary classification problem over a defined set of features that describes users based on their behaviour as well as the content they generate. Several works approached the identification of fake news spreaders or hate spreaders with diverse machine learning algorithms, such as logistic regression, random forest or SVM [1], [6], neural networks-based [2] and transformer models [3], [8].

The mentioned works explored a variety of features, including n-grams [4] or embeddings [5] for representing content, and lexical [6], psycholinguistic [7], personality [9], writing style [10] and emotions [11], for representing the user characteristics. For example, in [9] it was found that personality traits help to differentiate between fake news spreaders and fact-checkers and in [12] psycholinguistic features were used. The impact of visual features for profiling fake news spreaders was investigated in [13]. Unsupervised fact-finding algorithms that combine multi-modal microblog content features with analysis of propagation patterns to determine the veracity of microblog observations are presented in [14], considering also the possible presence of malicious sources. An explainable machine learning approach based on SHAP values was applied in [15] to understand the value of diverse features for predicting fake news spreaders.

Likewise, diverse features have been used for studies trying to characterize online hate spreaders. For example, in [16] the authors investigated the characteristics of accounts spreading hate, finding that they are different in activity, network centrality, and the type of content they produce. In [17] authors were concerned with separating users that propagate hate from those who counteract it, based on a lexical and psycholinguistic analysis of the accounts. In [18] it was suggested that modeling of user priors may be highly informative and complementary to improve pure content-based models in detecting hate speech.

Other types of online harm spreaders have been studied with the same strategies. Rumor spreaders were considered in [19], showing that users with a lower followee/follower ratio are more probable to spark the rumor diffusion, while those with a higher ratio are the one keeping them alive.

In [20] propagators of conspiracy theories were analyzed and compared with anti-conspiracy propagators based on the psychological and linguistic characteristics of both. A shared task was carried out in [21] for identifying potential Twitter users that spread stereotypes using indirect speech such as irony. In this task, also a variety of features were considered by different participating teams, including n-grams, writing style, personality and emotions, embeddings and transformers, combined with traditional supervised learning approaches.

The mentioned works in the literature focused on the distinction of regular users from users causing some specific harm (fake news, hate, irony, rumors, or others). In contrast, in this work, features spanning multiple dimensions are studied in order to better understand and characterize online harm spreaders as a whole.

III. Feature Extraction

Features in multiple dimensions are extracted in this work to characterize different types of spreaders. A large set of features is considered in an attempt to cover different aspects of a user behaviour that can be used to identify spreaders as well as distinguish among different types. Then, supervised models can be trained using these features to detect specific types of spreaders, such as those propagating fake news or hate speech, among other possibilities.

Since the goal of this work is to jointly analyze a variety of spreaders, features are extracted for capturing multiple dimensions or aspects of user behaviour as well as the user-generated content in social media.

Social media texts are analyzed by leveraging both lexicons and available trained models to extract features belonging to each dimension. The used lexicons include the Linguistic Inquiry and Word Count (LIWC) [22], NRC Emotion Lexicon [23], also known a *EmoLex*, and *Empath* [24]. Each of them allows to associate words to certain categories, such as topics or emotions.

The dimensions defined to characterized spreaders are the following:

- *Basic tweeting activity features* (5 features): features describing tweeting activity and the presence of certain elements in the tweets. These features include: whether the tweet is a retweet and the numbers of hashtags, mentions, URLs and emojis employed.
- *Stylometric features* (79 features): this group of features tries to capture the ways in which individuals express themselves and is analyzed at different levels: lexical (30 features), syntactical (42 features) and semantic (7 features). Among lexical features are character-level and word-level ones, such as total words, characters per word, frequency of large and unique words. Syntactic features denote the use of the most common part-of-speech (POS) tags, such as verbs, nouns, etc. Semantic features, try to capture meaning behind words, relates to categories like uncertainty and tentativeness, or the presence of certain cognitive processes.
- *Emotion-related features* (57 features): this group of features includes positive and negative emotions gleaned with different lexicons. For example, the eight basic emotions from *EmoLex* (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), a similar number from LIWC and a wider spectrum provided by *Empath*.
- *Sentiment-related features* (6 features): positive, negative and neutral sentiment estimated by features is extracted by leveraging both lexicons and trained models. VADER (Valence Aware Dictionary and sEntiment Reasoner) [25] provides sentiment-related scores. Stanza[1] NLP library provides a single sentiment score, and the the Emoji Sentiment Ranking [26] offers the sentiment conveyed by emojis.
- *Readability features* (9 features): features denoting the readability of texts produced by a user are based on metrics that estimate their difficulty in terms of the level of instruction needed to understand it. A number of readability formulas are used to quantify this aspect, including Flesch-Kincaid Grade Level, Fog Scale, SMOG Index, Automated Readability Index and Readability Consensus, among others.
- *Lexical richness features* (15 features): features indicating lexical richness of users measures the diversity of the vocabulary used by them. These features are also determined by formulas such as type-token ratio (TTR), Measure of Textual Lexical Diversity (MTLD), Hypergeometric distribution diversity (HDD), Yule's K, and others.

IV. Data description

Multiple shared tasks have proposed in events such as PAN@CLEF[2] for profiling different types of spreaders, releasing the corresponding datasets. Data of the three last editions of this event is used in this work for evaluation. PAN@CLEF 2020 Profiling Fake News Spreaders on Twitter task[3] [27] contains 100 tweets of 300 users labeled as fake news spreader o not. Likewise, the PAN@CLEF 2021 Profiling Hate Speech Spreaders on Twitter task[4] [28] contains 200 users identified as hate speech spreaders or not. Ultimately, PAN@CLEF 2022 Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO)[5] [21] task provided 200 tweets of 420 users, half of which spread irony, mostly towards stereotypes.

V. Experimental Results

In order to assess the relationship of the different kinds of features with the spreader class in each dataset, depending on their type, mutual information (MI) is calculated to observe the dependency between the variables. MI is equal to zero if and only if two random variables are independent, whereas higher values mean higher dependency.

[1]https://stanfordnlp.github.io/stanza/
[2]https://pan.webis.de/shared-tasks.html
[3]https://pan.webis.de/clef20/pan20-web/author-profiling.html
[4]https://pan.webis.de/clef21/pan21-web/author-profiling.html
[5]https://pan.webis.de/clef22/pan22-web/author-profiling.html

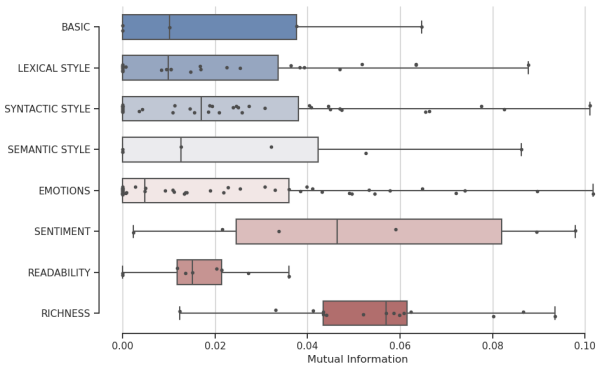
Figure 1 shows the MI values of all features with respect to the spreader class organized by the dimension of analysis in the three datasets. Features with high MI are assumed to be the most valuable for classifying spreaders, whereas null mutual information indicates features that can be removed before training as they would not be useful for prediction. Mean values of mutual information of each dimension allows to gain some insight of the overall influence of the dimension for characterizing the type of spreader in the dataset.

Observing the distributions of each dimension of features regarding fake news, hate speech and irony spreaders, figures (a), (b) and (c) respectively, it can be seen that each of them have the potential of impacting differently in the identification of spreaders. For example, sentiment-related features are, on average, more useful to identify fake news spreaders than hate speech spreaders or irony spreaders. Lexical features, instead, are more important for hate spreaders and even more for irony spreaders.

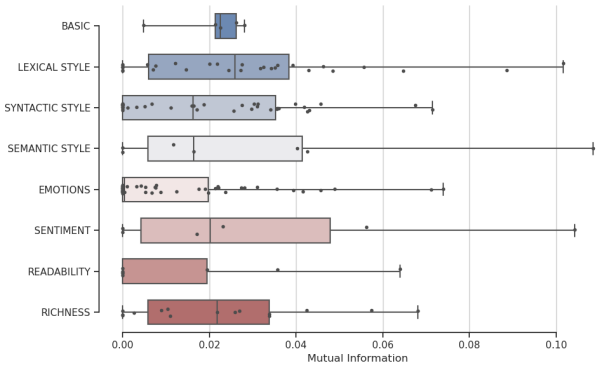
In all of the dimensions are outliers that have high values in terms of mutual information and are the ones better describing spreaders of a certain type as they separate them from non-spreaders. To illustrate this issue, the 57 emotion-related features can be further analyzed. The distributions showed many emotions having a MI equal to zero, but also many outliers, i.e. features with values largely above the average. The last features are the ones better separating spreaders from non-spreaders and, thereby, denoting the distinctive characteristics of a given type spreader. The features playing this role are different in each dataset as expected.

For the sake of comparison, Figure 2(a) to (c) summarize the top-10 features ordered by mutual information in the three datasets. Features in blue are those for which the average value of the feature is higher for non-spreaders, whereas in magenta are features with average values higher for spreaders. In the dataset of fake news spreaders, emotions like anticipation, positive, joy, cheerfulness are important, having higher values for non-spreaders. Also, fear, negative emotion and emotional, are important for classification according to MI, but having high values for spreaders. In the hate speech dataset, in turn, anger is the more informative emotion, and the top-10 is mostly occupied by negative emotions. Oppositely, non-spreaders of irony are the ones showing more emotions in general terms, either positive or negative, possibly denoting the ironic users intention of concealing their ironic comments.

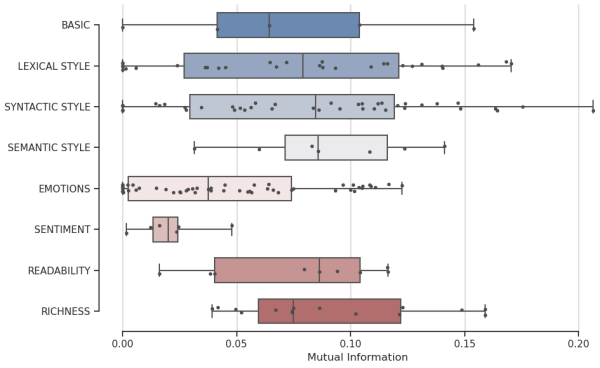
Finally, Table I shows the classification results for the three datasets using classical machine learning algorithms trained with the complete set of features. The results corresponds to the F1 scores and its standard deviation achieved using 10-fold-cross validation with default setting for each algorithm parameters. Binary classifiers were trained with each dataset for distinguishing between spreaders and non-spreaders. It is important to notice that the wide set of features used allows to classify the three types of spreaders with relatively good performance. Naturally, these results can be improved by tuning parameters or employing more sophisticated algorithms.



(a) Fake news spreaders



(b) Hate speech spreaders



(c) Irony spreaders

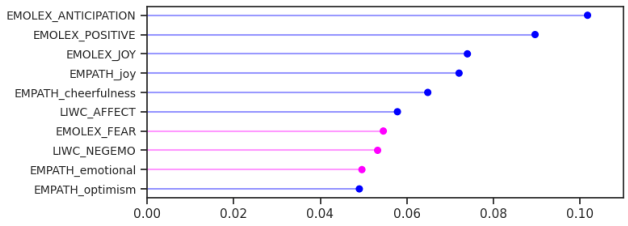
Fig. 1: Mutual information values for features in each dimension for the three datasets.

VI. Conclusions

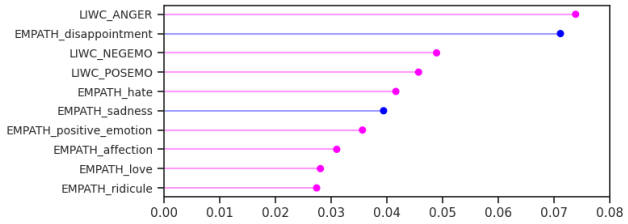
This work explored features grouped in multiple dimensions that allows to characterize the behaviour of multiple online harm spreaders. This broad set of features enables the identification of multiple types of spreaders, including fake news, hate speech and irony spreaders. Moreover, using these features it is possible to analyze the impact of each dimension in describing the behaviour of these users online. In future work, the similarities and differences of spreaders in terms of the defined dimensions will be explored.

TABLE I: F-measure classification results for spreaders.

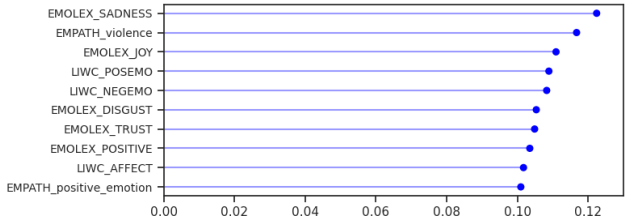
Classifier	Fake news spreaders	Hate spreaders	Irony spreaders
k-NN	63.08 ± 7.80	54.65 ± 8.86	82.33 ± 3.67
SVM (RBF)	71.68 ± 8.38	65.50 ± 14.55	94.80 ± 2.81
Random Forest	$\mathbf{74.77 \pm 11.69}$	$\mathbf{69.65 \pm 7.35}$	95.12 ± 3.63
Naïve Bayes	71.86 ± 5.82	62.70 ± 11.78	86.46 ± 6.30
XGBClassifier	71.15 ± 6.97	68.20 ± 8.77	$\mathbf{96.28 \pm 3.61}$



(a) Fake news spreaders



(b) Hate speech spreaders



(c) Irony spreaders

Fig. 2: Top-10 emotion-related features sorted by mutual information in the three datasets.

ACKNOWLEDGMENTS

This work has been partially funded by ANPCyT (Argentina) under grant PICT-2020-SERIEA-01375.

REFERENCES

[1] J. Buda and F. Bolonyai, "An ensemble model using n-grams and statistical features to identify fake news spreaders on Twitter–notebook for PAN at CLEF 2020," in *CLEF 2020 Labs and Workshops, Notebook Papers*, 2020.

[2] I. Vogel and M. Meghana, "Fake news spreader detection on Twitter using character N-grams," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, vol. 2696, Thessaloniki, Greece, 2020.

[3] S.-H. Wu and S.-L. Chien, "A BERT based two-stage fake news spreader profiling system," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, vol. 2696, Thessaloniki, Greece, 2020.

[4] I. Vogel and M. Meghana, "Detecting fake news spreaders on Twitter from a multilingual perspective," in *Proceedings of the 7th International Conference on Data Science and Advanced Analytics (DSAA 2020)*, 2020, pp. 599–606.

[5] H. B. Giglou, T. Rahgooy, J. Razmara, M. Rahgouy, and Z. Rahgooy, "Profiling haters on Twitter using statistical and contextualized embeddings," in *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, vol. 2936, 2021, pp. 1813–1821.

[6] H. B. Giglou, J. Razmara, M. Rahgouy, and M. Sanaei, "LSACoNet: A combination of lexical and conceptual features for analysis of fake news spreaders on Twitter," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, vol. 2696, Thessaloniki, Greece, 2020.

[7] H. R. M. Bello, L. Heilmann, and E. Ronan, "Detecting fake news spreaders with behavioural, lexical and psycholinguistic features," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, vol. 2696, 2020.

[8] A. Baruah, K. Das, F. Barbhuiya, and K. Dey, "Automatic detection of fake news spreaders using BERT," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, vol. 2696, Thessaloniki, Greece, 2020.

[9] A. Giachanou, E. A. Ríssola, B. Ghanem, F. Crestani, and P. Rosso, "The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers," in *Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB 2020)*, 2020, pp. 181–192.

[10] R. Labadie-Tamayo, D. Castro-Castro, and R. Ortega-Bueno, "Fusing stylistic features with deep-learning methods for profiling fake news spreaders," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, vol. 2696, Thessaloniki, Greece, 2020.

[11] E. Fersini, J. Armanini, and M. D'Intorni, "Profiling fake news spreaders: Stylometry, personality, emotions and embeddings," in *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, ser. CEUR Workshop Proceedings, vol. 2696, Thessaloniki, Greece, 2020.

[12] A. Giachanou, B. Ghanem, E. A. Ríssola, P. Rosso, F. Crestani, and D. Oberski, "The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers," *Data & Knowledge Engineering*, vol. 138, p. 101960.

[13] R. Cervero, P. Rosso, and G. Pasi, "Profiling fake news spreaders: Personality and visual information matter," in *Proceedings fo the International Conference on Applications of Natural Language to Information Systems (NLDB 2021)*, 2021, pp. 355–363.

[14] H. Shao, S. Yao, Y. Zhao, L. Su, Z. Wang, D. Liu, S. Liu, L. Kaplan, and T. Abdelzaher, "Unsupervised fact-finding with multi-modal data in social sensing," in *Proceedings of the 22th International Conference on Information Fusion (FUSION 2019)*, Ottawa, Canada, 2019, pp. 1–8.

[15] A. Shrestha and F. Spezzano, "Characterizing and predicting fake news spreaders in social networks," *International Journal of Data Science and Analytics*, vol. 13, pp. 385–398, 2022.

[16] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. M. Jr., "Characterizing and detecting hateful users on Twitter," *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM 2018)*, pp. 676–679, 2018.

[17] B. Mathew, N. Kumar, P. Goyal, and A. Mukherjee, "Interaction dynamics between hate and counter users on Twitter," in *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, Hyderabad, India, 2020, pp. 116–124.

[18] P. Chaudhry and M. Lease, "You are what you tweet: Profiling users by past tweets to improve hate speech detection," in *Information for a*

Better World: Shaping the Global Future: 17th International Conference, IConference 2022, 2022, pp. 195–203.

[19] A. Bodaghi and J. Oliveira, "The characteristics of rumor spreaders on Twitter: A quantitative analysis onreal data," *Computer Communications*, vol. 160, pp. 674–687, 2020.

[20] A. Giachanou, B. Ghanem, and P. Rosso, "Detection of conspiracy propagators using psycho-linguistic characteristics," *Journal of Information Science*, vol. 49, no. 1, pp. 3–17.

[21] R. O. Bueno, B. Chulvi, F. Rangel, P. Rosso, and E. Fersini, "Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO). Overview for PAN at CLEF 2022," in *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, 2022, pp. 2314–2343.

[22] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic inquiry and word count: LIWC 2001*. Lawrence Erlbaum Associates, Mahway, 2001.

[23] S. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[24] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, San Jose, California, USA, 2016, pp. 4647–4657.

[25] C. J. Gilbert and E. Hutto, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14)*, 2014.

[26] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of Emojis," *PLOS ONE*, vol. 10, no. 12, pp. 1–22, 2015.

[27] F. Rangel, A. Giachanou, B. Ghanem, and P. Rosso, "Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter," in *CLEF 2020 Labs and Workshops, Notebook Papers*, 2020.

[28] F. Rangel, B. Chulvi, G. L. D. L. Pena, E. Fersini, and P. Rosso, "Profiling hate speech spreaders on Twitter," Mar. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4603578