

Use of Lexical and Psycho-Emotional Information to Detect Hate Speech Spreaders on Twitter

Notebook for PAN at CLEF 2021

Riccardo Cervero

Università degli Studi di Milano-Bicocca (UNIMIB), Milan, Italy

Abstract

This notebook summarises the participation at the "Profiling Hate Speech Spreaders on Twitter" shared task [1] at PAN at CLEF 2021 [2], and describes the proposed method for the goal of binary classification into hate speech spreaders and non spreaders. This method consists in an ensemble method inspired by Buda-Bolonyai's previous work - based on the separate training of different baselines and the subsequent definition of a meta-model for the final prediction - has been proposed for both the English and Spanish corpora, with the introduction of more in-depth features relating to the personality traits of the users and the psychological and emotional dimensions detectable from the text they published. The aforementioned system achieved an accuracy result of 0.7 for the English-writing users' dataset and 0.8 for the Spanish-writing users' dataset (with a final average result of 0.75).

Keywords

Hate speech detection, personality traits, psycho-linguistic patterns, sentiment analysis

1. Introduction

The structure of online social networks, while able to offer several advantages - including the possibility of sharing content with thousands of users with ease - also encourages the proliferation of toxic narratives. In particular, the information filtering systems used to personalise the experience of each user have caused the intellectual isolation of certain sub-communities, called *echo chambers* [3]. These virtual bubbles are configured as closed virtual environments, and are extremely attractive to some readers because they propose tendentious contents that provoke a strong emotional engagement and because they exploit the so-called *confirmation bias* [4]. These ideological frameworks are the source of most of the hate messages spreading on the Web [5]. Given the scale of the phenomenon, it is necessary to exploit computational linguistics tools to stem its spread. This is precisely where the "Profiling Hate Speech Spreaders on Twitter" task at PAN at CLEF 2021 lies: the objective is to identify Twitter users who show a tendency to publish posts containing hate speech ("hate speech spreaders"), checking whether this can be done by extracting linguistic features from the last 200 tweets in their timeline. The task is carried out on two corpora in English and Spanish.


CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ r.cervero@campus.unimib.it (R. Cervero)

🆔 0000-0002-6642-9147 (R. Cervero)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In particular, this project will use features that can infer, from the raw text, the users' personality traits, the psychological and cognitive processes underlying the textual content and the emotional dimension of the published message. The underlying hypothesis is that individuals' psychological inclinations may influence not only their real-life interactions, but also their behaviour within the virtual community, as suggested by numerous studies [7, 8, 9]. For instance, it is clear from this research that trolling is correlated with mental disorders that often result in the dissemination of violent content: anti-social tendency, aggressive behaviour, psychopathy, narcissism and even the sadistic personality disorder (SPD).

The paper is organised as follows: Section 2 presents related works about the hate speech detection; Section 3 introduces the dataset provided at the task and describes the proposed system - outlining not only the predictive architectures employed, but also the features on which those has been trained. Lastly, the results and conclusions are discussed in Sections 4 and 5 respectively.

2. Related Work

The phenomenon of hate speech has recently become a popular area of research. The possibility of counteracting the dissemination of aggressive narratives towards specific targets by means of automatic processes has very often been tested. In general, a data-driven approach, i.e. extracting different types of features and exploiting them in combination with Machine Learning techniques to estimate a model that can produce the smallest possible error on new data, appears to be more frequent and effective. Classifiers used for this purpose are of various kinds: Naïve Bayes, as implemented by Kwok and Wang [10] in combination with a Bag-of-Words approach; Support Vector Machines, again applied on Bag-of-Words features by Greevy et al. [11]; Logistic Regression, trained, for instance, on N-grams, as it is the case with Waseem and Hovy's work [12] on Twitter users. Other more in-depth research has focused on identifying sub-classes of hate speech: Salminen et al. [13] have even developed a taxonomy of hate content in online social media, including the four main macro-categories of *accusation*, *humiliation*, *swearing* and *promotion of violence*. As the tools applied to Natural Language Processing evolved, more complex approaches were tested in various studies. One case is that of the introduction of Deep Learning techniques: Mikolov et al. [14] use embeddings as features; in other many studies, one can find deep architectures such as Convolutional Neural Networks [15], Recurrent Neural Networks [16], a combination of both [17], transformers - in particular, BERT [18]. In general, however, the best results are often offered by ensemble methods. One case is that of MacAvaney et al. [19], who have exploited the innovative multi-view learning strategy, creating separate *view-classifiers* for groups of different features and then combining them with a Linear Support Vector Machine to produce a meta-model. Other cases are [20] and [21]. Given the effectiveness of the latter solution, in this project an ensemble method for profiling hate speech spreaders has been chosen.

3. Method

This Section will illustrate the proposed system for identifying Twitter users who spread hate messages, based on the linguistic and semantic features that can be extracted from the texts they published in the past.

Software submission was made via TIRA platform [22].

3.1. Datasets

The two datasets provided for the "Profiling Hate Speech Spreaders on Twitter" task are both composed of 200 observations, each recording in a unique sub-corpus the last 200 tweets published by the respective anonymous user included in the sample - whose username was converted into an alphanumeric sequence by means of a hashing algorithm. No other operations were performed on the original text, except the replacement with specific tokens of any URLs, hashtags and names of the users mentioned or retweeted.

3.2. Environment Setup

The entire project was developed within the Python programming environment, exploiting version 3.7. The main libraries used are pandas¹ and numpy² for the management of basic data structures, scikit-learn³ for the application of the key Machine Learning techniques - such as baselines' training and their validation -, and xgboost⁴ for the implementation of the gradient boosting method.

3.3. Ensemble Model

The architecture of the proposed predictive model is inspired by the work of Buda & Bolonyai [6] at the "Profiling Fake News Spreaders on Twitter" task at PAN at CLEF 2020, aimed to inquire the feasibility of detecting authors who had shared fake news in their past timeline, only looking at the linguistic features extractable from the posts they had published. Their system achieved the best overall accuracy on the English corpus (0.75), and has been tied for first place as far as the average result on both datasets (0.775). In detail, after having trained 4 baselines (a regularized Logistic Regression, one Support Vector Classifier, a Random Forest method and an implementation of the gradient boosting algorithm offered by the XGBoost library) on N-grams collected from the sub-corpora and having estimated a fifth model by implementing again the gradient boosting method on some stylistic features (which will be described later in the Section 3.4.4), a Logistic Regression it is applied as a meta-model that estimated the relative weights of the predictions made separately by each of the stacked baselines, and that in turn returned a final binary prediction about the tendency of the user to spread false information or not. Unlike done with the model originally presented by the authors, in this case it was decided to also

¹Official documentation at <https://pandas.pydata.org>.

²Official documentation at <https://numpy.org/>.

³Official documentation at <https://scikit-learn.org/>.

⁴Official documentation at <https://xgboost.readthedocs.io/>.

experiment with a Ridge Classifier as a meta-model, selecting *ex post* the best solution on the basis of the accuracy result produced.

The first four baselines undergo a training process consisting in an extensive grid search of the optimal combination among two text pre-processing methods, different vectorization techniques and the parameters and hyperparameters of the models themselves. More specifically, as far as the pre-processing stage, both pipelines convert all the tokens to lower case and remove non alphanumeric characters; the only difference is that the second pipeline implemented preserves the emoticons and emojis. With regard to the corpus vectorization, the Term Frequency – Inverse Document Frequency function is applied to different ranges of N-grams, considering unigrams, bigrams and both, as well as running tests to optimise the hyper-parameter value which set their minimum overall document frequency. Therefore, in conjunction with the search for the best pre-processing and vectorization strategies, the sub-optimal parameters of the four baselines have been found by using a Grid Search Cross Validation technique, with the number of folders set at 5 by the original authors and 10 in the case of this project. As a last step, in order to prevent the ensemble model from overfitting the training set, the Logistic and Ridge Regression have not been trained directly on the predictions of the baselines, but on the approximation of the predictions distribution, obtained by refitting the sub-models with the cross-validated hyperparameters on different chunks of the training set.

Thus, as has just been explained, from both corpora provided for the "Profiling Hate Speech Spreaders on Twitter" task at PAN 2021 N-grams were collected on which 4 sub-models were trained respectively; these were then stacked together with a fifth baseline: an XGBoost algorithm, whose input consists of a different category of features (in Buda and Bolonyai's original work composed mostly of some stylistic statistics). The main contribution of this paper lies in the combination of the aforementioned approach by Buda & Bolonyai, which has proved effective in profiling fake news spreaders at PAN 2020 shared task, and a set of features capable of synthesising the personality of the authors and bringing out certain psychological and emotional dynamics that can be useful in accurately classifying Twitter users into 'hate speech spreaders'. The next Section will explain these features in more detail.

3.4. Features

Therefore, the new ensemble model includes a fifth baseline, a gradient boosting algorithm, which receives new feature sets as input. In this project, we tested all the possible combinations between the original stylistic features proposed by Buda and Bolonyai (Section 3.4.4) and the different features related to personality traits (Section 3.4.1), psycho-linguistic patterns (Section 3.4.2) and emotions and polarity (Section 3.4.3) extrapolated from the text. In the end, the best results were respectively offered: for the English corpus, by the mix of personality-related features with emotional and sentiment dimensions tagged within the text; for the Spanish corpus, again by the combination of the personality scores extracted through the Five Factor Model, together with psycho-linguistic patterns found by LIWC software, the stylistics features extracted by Buda and Bolonyai, and still the emotions. These respective solutions have been therefore proposed for the computation of the final accuracy on the test set held by the event organisers.

It is important to note that, in both optimal feature sets, the features describing authors'

personality traits are present: this demonstrates the validity of the initial hypothesis, according to which the identification of users who publish violent content can be reasonably conducted through a psychological profile performed with computational tools.

3.4.1. Five Factor Model Features

The Five Factor Model [23] (FFM) consists in a process of attributing certain psychological characteristics to an individual according to the so-called 'Big Five' taxonomy, developed by Rothmann and Coetzer [24] as a modern evolution of the dispositional approach to the study of human personality and its consequences on behaviour. This theory - also 'trait theory' - stems from the discovery of semantic associations as a result of statistical analysis carried out on a sample of personality survey data. Thanks to these evidences, it was possible to demonstrate that the human psychological dimension can be summarised in only five aspects, referred to by words and expressions recurrent in natural language during the description of the personality of an individual. The five suggested standard factors are listed below: openness to experience, conscientiousness, agreeableness, extraversion, and emotional stability. Hence, this approach argues for the existence of semantic associations between different sets of words and each of the five factors through which it appears possible to synthesise the personality of an individual. From this theoretical basis, Neuman and Cohen's method [25] derives a vector of personality scores which can be provided as input to predictive Machine Learning models. In details, these personality scores are calculated by computing the cosine similarity between the context-free embedding representations of both the input text - written by the selected author - and a set of benchmark adjectives empirically observed as to be able to encode the essence of personality (and selected by Neuman and Cohen).

3.4.2. LIWC Features

The Linguistic Inquiry and Word Count (LIWC) [26] is a software developed for Natural Language Processing tasks, which is able to automatically detect linguistic patterns and map the text into a dense representation composed of 73 psychologically-meaningful linguistic categories. Therefore, this strategy is configured as a lexicon-based method that associates, within a dictionary, a set of predefined classes to several tokens. Using this dictionary, it is then possible to tag the sought psycho-linguistic features and thus obtain evidence of mental and cognitive processes underlying the text of the tweets. The resource used in this project is the LIWC2015 dictionary⁵, for both languages. This tool considers several inflected variants from about 6400 word stems, as well as certain selected emoticons; each of these linguistic items has been assigned one or multiple categories among the mentioned 73. These psycho-linguistic classes are arranged in a hierarchical structure: three macro-categories have been set up by the creators of the software, which are in turn divided into numerous tags. The three macro-categories with which psycho-linguistic patterns can be labelled are: linguistic dimensions - comprising different types of function words, such as pronouns, article, prepositions, etc.; grammars - including common verbs and adjectives (like *eat, come, free, happy, long*), comparisons (*greater, best, after*), interrogatives (*how, when, what*), numbers, etc.; psychological processes - for instance, of an

⁵Official website of LIWC software is at: <https://liwc.wpengine.com>.

affective, cognitive, social or perceptual nature. In conclusion, the LIWC features, with which the Machine Learning models will be trained, coincide with a 73-dimensional vector indicating the total raw occurrence of the mentioned sub-categories within the sub-corpora associated with each author.

3.4.3. Emotional Features

The aforementioned components have been combined with a vector recording the raw occurrence of eight emotional dimensions tagged within the text by exploiting the association proposed by the NRC lexicon⁶. In the same way, two dimensions related to the sentiment polarization have been extracted.

3.4.4. Original Features

Finally, the potential of the stylistic features proposed by Buda and Bolonyai (originally used for the task of fake news spreaders detection in Twitter) has also been tested on the new task of profiling hate speech spreaders, both in combination with those presented in the previous sections and alone. In details, this type of feature is configured as a set of user-wise statistics: (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.

4. Results

In this Section, we will present the best results obtained from testing the several new feature sets - derived from all the possible combination of the features presented in Section 3.4 - on which the fifth baseline is trained. The main criterion for selecting the best models on the English and Spanish corpora - then proposed to the "Profiling Hate Speech Spreaders on Twitter" task at PAN 2021 - coincides with the maximisation of the average accuracy obtained on the 10 folds derived from the split of the training set during the Cross Validation procedure, and, in case of very similar results, also with the minimisation of the variance of these accuracy values. As visible in Table 1, the best result on the English corpus is produced by combining Five Factor Model features with lexicon-based emotional and sentiment dimensions. As far as the Spanish dataset, the best performing solution consist in the union of FFM personality scores, LIWC features, the eight emotional dimensions and the Buda-Boloyai's original set. The first system provided an accuracy of 0.7 on the English test set; the second achieved a result equal to 0.8 on the Spanish test set.

These results underline the validity of the initial hypothesis - that the extraction of personality characteristics is extremely useful for the binary classification of users into hate speech spreaders and non spreader -, and the validity of an ensemble approach for the same task.

⁶Official NRC Lexicon is at: <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>.

Table 1

Best overall solution for both English and Spanish corpora.

Language	Features	Accuracy (Train Set)	Accuracy (Test Set)
EN	FFM + Emo. + Sent.	0.695	0.7
ES	FFM+ LIWC+ Emo.+ Styl.	0.845	0.8

5. Conclusion

The ease and speed with which today's Web technologies allow information to be shared has made online social media an extremely dangerous and effective means for disseminating offensive messages, raising the need for automated tools that can stop the flow of toxic information before it contaminates the virtual community. In this paper, hate speech detection is approached from an author profiling perspective: instead of analyzing the single content, the aim is to identify users who tend to publish posts that fall into the category of "hate speech". For the participation to the "Profiling Hate Speech Spreaders on Twitter" task at PAN at CLEF 2021, an ensemble method inspired by a previous work by Buda and Bolonyai is proposed for the detection of fake news spreaders: four baselines are trained on N-grams, and a fifth one receives as input features defined by them. The main contribution of this paper is to propose as features of the fifth baseline, instead of the descriptive statistics related to writing style originally employed by Buda and Bolonyai, a set of features related to personality traits (defined by the Five Factor Model), and representing psycho-linguistic patterns and emotional dimensions from the text. Thanks to these strategies, the result obtained on the test set provided by the task organisers at PAN 2021 is 0.7 and 0.8 for the English and Spanish corpora respectively.

References

- [1] Rangel F., Liz De La Peña Sarracén G., Chulvi B., Fersini E., Rosso P. "Profiling Hate Speech Spreaders on Twitter Task at PAN 2021". In: Faggioli G., Ferro N., Joly A., Maistro M., Piroi F. "CLEF 2021 Labs and Workshops, Notebook Papers" Conference and Labs of the Evaluation Forum (CLEF 2021), CEUR-WS.org (2021).
- [2] Bevendorff J., Chulvi B., Liz De La Peña Sarracén G., Kestemont M., Manjavacas E., Markov I., Mayerl M., Potthast M., Rangel F., Rosso P., Stamatatos E., Stein B., Wiegmann M., Wolska M., Zangerle E. "Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection". In: Selcuk Candan K., Ionescu B., Goeuriot L., Larsen B., Müller H., Joly A., Maistro M., Piroi F., Faggioli G., Ferro N. "12th International Conference of the CLEF Association (CLEF 2021)", Springer, Bucharest, Romania (2021).
- [3] Sunstein C. R. "The law of group polarization". In: *Journal of political philosophy* 10 (2002), pp. 175–195.
- [4] Nickerson R. S. "Confirmation bias: A ubiquitous phenomenon in many guises." In: *Review of general psychology* 2(2) (1998), pp. 175.

- [5] Del Vicario M. et al. "Echo chambers: Emotional Contagion and Group Polarization on Facebook" In: *Proceedings of the National Academy of Sciences* 113(3) (2016), pp. 554–559.
- [6] 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., Névél A. (eds.) *CLEF 2020 Labs and Workshops, Notebook Papers*, CEUR-WS.org (2020)
- [7] Andjelovic T., Buckels E. E., Paulhus D. L., Trapnell P. D. "Internet trolling and everyday sadism: Parallel effects on pain perception and moral judgment". In: *Journal of Personality* 87(2) (2019), pp. 328–340.
- [8] Buckels E. E. "Probing the Sadistic Minds of Internet Trolls". In: *Society for Personality and Social Psychology* (2019).
- [9] March E., Steele G. "High Esteem and Hurting Others Online: Trait Sadism Moderates the Relationship Between Self-Esteem and Internet Trolling". In: *Cyberpsychology, Behavior, and Social Networking* 23(7) (2020), pp. 441–446.
- [10] Kwok I, Wang Y. "Locate the hate: Detecting tweets against blacks". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence AAAI'13* (2013), pp. 1621–1622.
- [11] Greevy E, Smeaton AF. "Classifying racist texts using a support vector machine". In: *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR '04* (2004), pp. 468–469.
- [12] Waseem Z, Hovy D. "Hateful symbols or hateful people? Predictive features for hate speech detection on twitter". In: *Proceedings of the NAACL Student Research Workshop* (2016), pp. 88–93.
- [13] Salminen J., Almerikhi H., Milenkovic M., Jung S., Kwak H., Jansen B.J. "Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media" (2018).
- [14] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. "Distributed representations of words and phrases and their compositionality". In: *Proc. NIPS* (2013), pp. 3111–3119.
- [15] Badjatiya P., Gupta S., Gupta M., Varma V. "Deep learning for hate speech detection in tweets". In: *Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion* (2017), pp. 759–760.
- [16] Del Vigna F., Cimino A., Dell'Orletta F., Petrocchi M., Tesconi M. "Hate me, hate me not: Hate speech detection on facebook". In: *ITASEC* (2017).
- [17] Huynh T. V., Nguyen V. D., Nguyen K. V., Nguyen N. L. T., Nguyen A. G. T. "Hate speech detection on vietnamese social media text using the bi-gru-lstm-cnn model" (2019).
- [18] Devlin J., Chang M., Lee K., Toutanova K. "BERT: pre-training of deep bidirectional transformers for language understanding" (2018).
- [19] MacAvaney S., Yao H. R., Yang E., Russell K., Goharian N., Frieder O. "Hate speech detection: Challenges and solutions". *PLoS ONE*. 14(8):1–16 (2019).
- [20] Nina-Alcocer V. "Vito at HASOC 2019: Detecting hate speech and offensive content through ensembles". In: Mehta P., Rosso P., Majumder P., Mitra M. (eds.) *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, Kolkata, India, December 12-15, 2019, *CEUR Workshop Proceedings*, vol. 2517, pp. 214–220. CEUR-WS.org (2019).
- [21] Nourbakhsh A., Vermeer F., Wiltvank G., Van der Goot R. "Struggle at SemEval-2019 task 5: An ensemble approach to hate speech detection." In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 484–488. Association for Computational Linguistics,

Minneapolis, Minnesota, USA (2019).

- [22] Potthast M., Gollub T., Wiegmann M., Stein B. "TIRA Integrated Research Architecture". In: Ferro N., Peters C. "TIRA Integrated Research Architecture" in "Information Retrieval Evaluation in a Changing World", Springer, Berlin Heidelberg New York, DOI: 10.1007/978-3-030-22948-1_5 (2019).
- [23] John O.P., Srivastava S. "The Big-five Trait Taxonomy: History, Measurement, and Theoretical Perspectives". In: Handbook of Personality: Theory and Research (1999), pp. 102–138.
- [24] Rothmann S., Coetzer E. P. "The big five personality dimensions and job performance". In: SA Journal of Industrial Psychology (29) (2003).
- [25] Neuman Y., Cohen Y. "A Vectorial Semantics Approach to Personality Assessment". In: Scientific Reports 4(1) (2014).
- [26] Pennebaker J. W., Boyd R. L., Jordan K., Blackburn K. "The Development and Psychometric Properties of LIWC 2015". Tech. rep (2015).