Phonetic Detection for Hate Speech Spreaders on Twitter

Notebook for PAN at CLEF 2021

Edwin Puertas[1], Juan Carlos Martinez-Santos[1]

[1]*Universidad Tecnológica de Bolívar, Cartagena, Colombia*

Abstract

Nowadays, hate messages have become the object of study on social media. Efficient and effective detection of hate profiles requires various scientific disciplines, such as computational linguistics and sociology. Here, we illustrate how we used lexical and phonetic features to determine if the author spreads hate speech. This article presents a novel strategy for the characterization of the Twitter profile based on the generation of lexical and phonetic user features that serve as input to a set of classifiers. The results are part of our participation in the PAN 2021 in the CLEF in the task of Profiling Hate Speech Spreaders on Twitter.

Keywords

phonetic syllable, phonetic feature, feature extraction, hate speech spreader

1. Introduction

Hate speech on social media is a complex phenomenon whose detection has recently gained significant traction in the natural language processing community, as several recent review papers attest. For this reason, knowing the profile of an author can be of vital importance on social media platforms because nowadays, hate messages on these networks have become more common. Lexical also plays an essential role in the development of hate speech detection systems [1]. Insecurity, to identify psychological traits that allow the detection of profiles with abnormal behaviors that may cause harm to other users or discover false profiles (a person can have multiple profiles for fraud and other misdeeds) [2, 3].

In the literature, the research aims to study the sources of micro-microblogging from the lexical, syntactic, and semantic levels in particular, excluding other levels of study, which are relevant when identifying the form of a word taking into account its structure, meaning, and phonetics. For this reason, it is necessary to understand the information in any context and interpret it automatically to detect features at the lexical, syntactic, morphological, and phonological levels of the texts that allow the representation of knowledge in natural language. For this reason, it is imperative to be able to analyze microblogging sources using new phonetic syllable

embedding models, which contribute as predictive features to determine the intentionality of a written text [4].

In this document, the proposal is described as part of our participation in the Profiling Hate Speech Spreaders on Twitter task PAN 2021 at CLEF [5, 6]. Also, the presentations were made on the TIRA [7] platform, in which we configure a virtual server with ten processors. This task focuses on investigating that, given a Twitter feed, determines if its author spreads hate speech. To do this, we study the generations and the analysis of different lexical and phonetic features to assess the hate profiles on Twitter.

The rest of the paper is structured as follows. In Section 2, we introduce the related work. Section 3 presents the details of the proposed strategy. In Section 4 and 5, we discuss the experiments and analysis of Results. We conclude in Section 6 with remarks and future work.

2. Related work

Hate speech is a language that attacks or diminishes, inciting violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or other. It can occur with different linguistic styles, even in subtle forms as humor [8]. The automatic detection of profiles of hate speech senders on social media has caught the attention of researchers in recent years. Many techniques are proposed in the literature to analyze this problem. Fortuna [8] show out a systematic review in the field where they frame the problem, its definition, and identify methods and resources. In addition, they take a systematic approach that critically analyzes theoretical aspects and practical resources, such as data sets and other projects.

Currently, some researchers address the issue of hatred as other issues. Caetano et al.[9] present in his work several alternative implementations of three of these tasks: hate speech, aggressive behavior, and recognition of the target group. They offered different learning methods, such as regularized logistic regression, convolutional neural networks (CNN), and deep bidirectional transformers (BERT). They also used word embeddings, word n-grams, character n-grams, and psycholinguistics-motivated characteristics (LIWC) alike. The results suggest that a purely data-driven BERT model, and to some extent also a hybrid CNN model with psycholinguistic information, generally outperforms the alternatives considered for all tasks in both English and Spanish.

On the other hand, in several international competitions such as the PAN in its different editions, they address similar problems, one of which is worth highlighting the characteristics used to detect some intention in social networks. Of the most used attributes in PAN, such as n-grams, stylistics, personality and emotions, and word embeddings or sequences embedding [10] [11]. In the same way, the International Workshop on Semantic Evaluation (SemEval) has proposed tasks that seek to address such approaches. One case is task 5 SemEval 2019 [12], which goal is detecting hate speech against immigrants and women in messages in Spanish and English taken from Twitter. The task is organized into two related classification subtasks: detecting hate speech and identifying other characteristics in hate content.

3. System Description

In this section, we describe the predictive model used in our submission. The model used for Profiling Hate Speech Spreaders on Twitter at PAN 2021 determines whether the author of a Twitter feed spreads hate speech. Following the data set's attributes and the task's objectives, we have proposed the following hypothesis. How are the lexical and phonetics features determining to determine hate profiles on Twitter?

Under the hypothesis presented, we proposed two strategies. The first generates lexical characteristics from terms, hashtags, mentions, URLs, and emojis. Additionally, metrics include avg words, kurtosis, skew words, the number of adverbs (neg, time, place, mode, quantity), adjectives, hate words, singular nouns, and plural nouns. We combined them with syntactic characteristics such as noun phrases, verbs, adjectives, and other types of expression. We also include the calculation of polarity using the Senticnet 5 approach[13].

The second strategy focuses on phonetic features through the syllabification process, which consists of calculating phonetic syllables. Similarly, to generate individual phoneme embeddings for each language. Finally, we calculate the frequencies of use of the phonemes for each of the languages in the training dataset. Based on the proposed strategies, we designed the Training System. Figure 1 shows the proposed system to determine if its author spreads hate speech, consisting of the following stages: data reading, data cleaning, generation of examples, extraction of characteristics, classification, and tests.
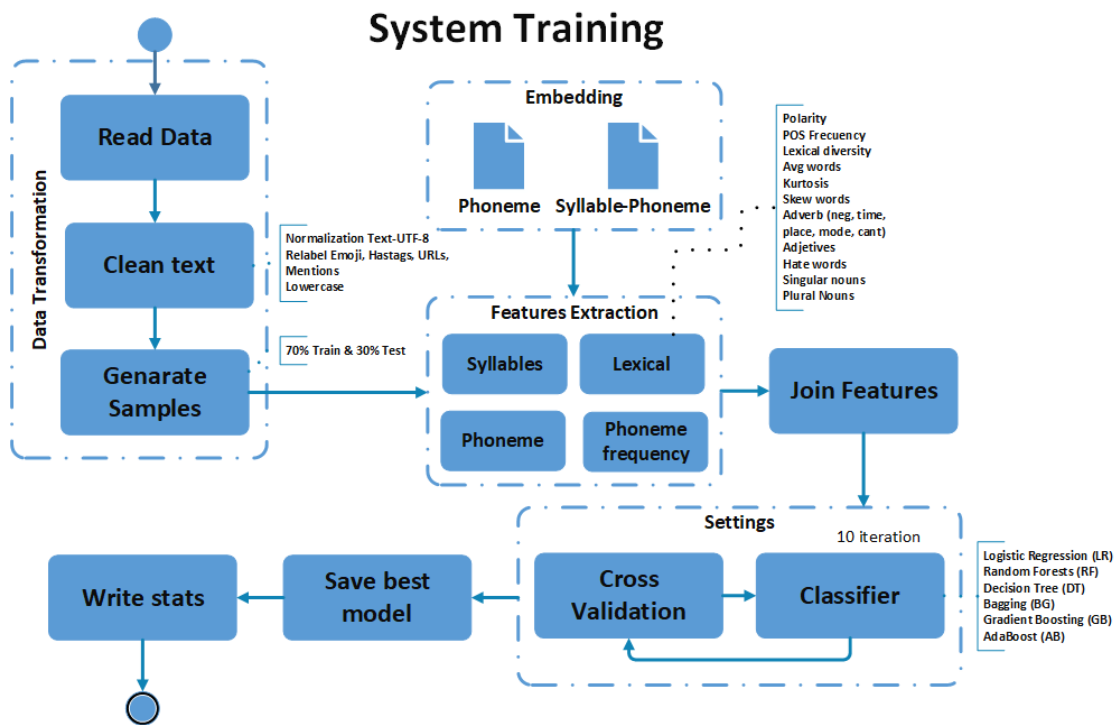

Figure 1: System Training

3.1. Data Description

Profiling Hate Speech Spreaders' data set on Twitter at PAN 2021 consists of 200 tweets in XML file per author, where the file's name corresponds to a unique author identifier. The set also includes texts in Spanish and English.

3.2. Embeddings

For the transformation of phonetic characters, phonetic syllable inlay models are used [14, 15] and phonemes. For the training of phonetic syllables and phonemes, we used the corpus CESS-ESP [16] for Spanish and BROWN corpora [17] for English, since they are sources that provide a standard syntactic and lexical structure for this study. Regarding the vector transformation, before the training of the inlays. First, we extract all the sentences from CESS-ESP and BROWN corpus, and we generated a vector of words for each sentence of the corpus. Second, we removed the stopwords from the word vectors. Subsequently, we extracted the syllables for each word in the vector. Then each syllable is transformed into its phonetic representation by IPA (International Phonetic Alphabet) using the transliterate function of the Epitran [18]. Concerning phoneme embeddings, for each character of the word, its IPA phonetic representation is carried out utilizing Epitran's trans_list function.

On the other hand, for the transformation of lexical characteristics, we determined the initial sentiment of the message using SenticNet 5 [13], which performs an analysis of expressions of several words that do not explicitly convey emotions but are related to concepts that do. Then, other lexical characteristics are determined, such as those described in [19].

3.3. Data Transformation

In this stage, we concatenate each user's tweets to have only one document per user profile. Then we clean up the texts by replacing the hashtags with the word "hashtag", the mentions with the word "mention", the URLs with the word "URL", and the emojis with the word "emoji". In addition, we remove special characters, punctuation marks except for semicolons, parentheses, and punctuated numbers. Then, the globally relabeled words are searched and counted. Finally, we divide the dataset into 70% for training and 30% for testing.

3.4. Feature Extraction

This phase focuses on extracting lexical and phonetic features. We divided the lexical features' extraction process into several activities: the use of words, hashtags, mentions, URLs, emojis, polarity, frequent use of POS, and hateful words.

Regarding the extraction of phonetic characteristics, we divided the process into three specific tasks: phonetic syllables, phonetic frequencies, and phonemes. About phonetic syllables, first, the embedded phonetic syllables trained are loaded into memory. Second, for each message, the sentences are extracted. Third, you get the words for each of the identified sentences. Fourth, we get syllables from each word. Fifth, we encoded syllables in IPA. Sixth, from the model of phonetic syllables inlays, the vector of characteristics of each phonetic syllable is obtained. Seventh, the feature vectors are averaged into a single vector.

For phonetic frequencies, first, the trained phoneme embeds are loaded into memory. Second, for each message, the sentences are extracted. Third, we represent each sentence in a list of phonemes using the function trans_list from the Epitran library. Fourth, we counted the phonemes of the sentence using the alphabet of the study language as a reference.

Regarding phonemes, first, the trained phonetic embeddings are loaded into memory. Second, for each message, the sentences are extracted. Third, we represented the sentence in a list of phonemes using the transliterate function of the Epitran library. Fourth, we obtained the sentence phonemes. Fifth, they are encoded in IPA. Sixth, from the model of phonetic syllables inlays, the vector of characteristics of each phonetic syllable is obtained. Seventh, the feature vectors are averaged into a single vector.

3.5. Settings

At the configuration stage, the system will adjust machine hardware parameters such as processors and threads. In addition, the system can configure different scenarios for the use of the classifiers. Finally, we may adjust the system to store the best-performing vector words and qualifiers. We should note the system used 70% of the data for training and the remainder 30% for testing during all the experiments.

On the other hand, based on the previous tasks carried out in the PAN, several classifiers were examined, such as Logistic Regression (LR), Random Forests (RF), Decision Tree (DT), Bagging (BG), Gradient Boosting (GB) and AdaBoost (AB).

4. Experiments and Analysis of Results

We carried out different experiments during the pre-evaluation phase, and we took the best ones into account for the evaluation phase. We used the usual competition metrics, including Accuracy (Acc), to evaluate our system. We will explain the best techniques to determine whether its author spreads hate speech in the pre-evaluation phase in detail in the following sections.

We should note that the system presented was trained and tested with the dataset provided by the official site of PAN 2021 [20]. Table 1 shows the results obtained after evaluating our system with the training dataset. The system uses various classification algorithms, such as Logistic Regression (LR), Random Forests (RF), Decision Tree (DT), Bagging (BG), Gradient Boosting (GB), and AdaBoost (AB). But in the case of the English language, AdaBoost obtained better performance. And for the Spanish language, Random Forest had better accuracy.

5. Result Test

Table 2 show the performance of the models using the training dataset, the test dataset. On the scale of the best results, our model occupied the 46th position.

Table 1
Summary of results in spreads hate speech classification per language

Classifiers	ES		EN	
	Acc	std	Acc	std
Logistic Regression	0.70	(+/- 0.064)	0.55	(+/- 0.041)
Random Forest	**0.73**	**(+/- 0.066)**	0.56	(+/- 0.053)
Decision Tree	0.63	(+/- 0.049)	0.50	(+/- 0.058)
Bagging	0.72	(+/- 0.059)	0.54	(+/- 0.073)
AdaBoost	0.70	(+/- 0.051)	0.57	(+/- 0.082)
Gradient Boosting	0.70	(+/- 0.041)	**0.60**	(+/- 0.052)

Table 2
Ranking of results in spreads hate speech classification per language

Lang	Dataset Training	Dataset Test
EN	0.60	0.60
ES	0.73	0.76
AVG	0.67	0.68

6. Discussion and Conclusion

The task of Profiling Hate Speech Spreaders on Twitter CLEF PAN 2021 [5] involved different tasks. The first task was preprocessing the corpus, composed of 200 tweets for each user profile, for 10,000 posts. Fortunately, quality assurance was not a challenge because the task committee cleaned the tweets and balanced the dataset for each target class. On the contrary, feature extraction was one of the most significant challenges because it was necessary to achieve good performance with few texts per user profile.

Additionally, the result evidenced that the experimental results fulfill the proposed hypothesis. Likewise, they showed that phonetic features are relevant when determining whether an author spreads hate speech on social networks. It is established in future works that it is necessary to explore a more elaborate classification using different feature extraction techniques such as Late Fusion, Importance of characteristics, and Elimination of recursive characteristics.

References

[1] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1–10.

[2] A. Ferrari, A. Consoli, Building accurate hav exploiting user profiling and sentiment analysis, ArXiv abs/1609.07302 (2016) 1–595.

[3] M. Fatima, K. Hasan, S. Anwar, R. M. A. Nawab, Multilingual author profiling on facebook, Inf. Process. Manage. 53 (2017) 886–904. URL: https://doi.org/10.1016/j.ipm.2017.03.005. doi:10.1016/j.ipm.2017.03.005.

[4] E. Puertas, J. A. Alvarado, Modelo que mejore la detección de polaridades hechas con word embedding con la ayuda de predictores fonéticos y el apoyo de elementos emocionales., in: ENEDI-2020, ENEDI-2020, https://www.acofi.edu.co/eiei2020/wp-content/uploads/2020/10/Memorias-ENEDI . . ., 2020, pp. 95–104.

[5] F. Rangel, P. Rosso, G. L. D. L. P. Sarracén, E. Fersini, B. Chulvi, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021, pp. 1–7.

[6] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection, in: 12th International Conference of the CLEF Association (CLEF 2021), Springer, 2021, pp. 1–7.

[7] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019, pp. 1–7. doi:10.1007/978-3-030-22948-1_5.

[8] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[9] S. Caetano da Silva, T. Castro Ferreira, R. M. Silva Ramos, I. Paraboni, Data driven and psycholinguistics motivated approaches to hate speech detection, Computación y Sistemas 24 (2020).

[10] F. Rangel, P. Rosso, Overview of the 7th author profiling task at pan 2019: bots and gender profiling in twitter, in: Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop, 2019, pp. 1–7.

[11] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter, in: CLEF, 2020, pp. 1–7.

[12] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.

[13] E. Cambria, S. Poria, D. Hazarika, K. Kwok, Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018, pp. 1–8.

[14] E. Puertas, Embedding of phonetic syllables in english, 2020. URL: https://doi.org/10.5281/zenodo.4299251. doi:10.5281/zenodo.4299251.

[15] E. Puertas, Embedding of phonetic syllables in spanish, 2020. URL: https://doi.org/10.5281/zenodo.4299242. doi:10.5281/zenodo.4299242.

[16] M. A. M. Antonín, M. T. Delor, L. Màrquez, M. Bertran, Anotación semiautomática con papeles temáticos de los corpus cess-ece, Procesamiento del Lenguaje Natural (2007) 67–76.

[17] C. Macleod, N. Ide, R. Grishman, The american national corpus: A standarized resource for american english., in: LREC, 2000, pp. 1–7.

[18] D. R. Mortensen, S. Dalmia, P. Littell, Epitran: Precision G2P for many languages, in: N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings

of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, France, 2018, pp. 1–4628.

[19] E. Puertas, L. G. Moreno-Sandoval, F. M. Plaza-del Arco, J. A. Alvarado-Valencia, A. Pomares-Quimbaya, L. Alfonso, Bots and gender profiling on twitter using sociolinguistic features, CLEF (Working Notes) (2019) 1–8.

[20] F. RANGEL, B. CHULVI, G. L. D. L. PEÑA, E. FERSINI, P. ROSSO, Profiling hate speech spreaders on twitter, 2021. URL: https://doi.org/10.5281/zenodo.4603578. doi:10.5281/zenodo.4603578.