Profiling Hate Speech Spreaders on Twitter: Transformers and mixed pooling

Notebook for PAN at CLEF 2021

Álvaro Huertas-García[1,2], Javier Huertas-Tato[1], Alejandro Martín[1] and David Camacho[1]

[1]*Department of Computer System Engineering, Universidad Politécnica de Madrid, Calle de Alan Turing, 28031, Madrid, Spain*

[2]*Department of Computer Sciences, Universidad Rey Juan Carlos, Calle Tulipán, 28933, Madrid, Spain*

Abstract

The growth of Online Social Networks (OSNs) has originated an increasing presence of online hate-spreaders. This fact undermines the integrity of online conversations by sharing inflammatory claims that influence public opinion while sow conflict on social or political issues. In this work, we propose a system for Authors Profiling Hate Speech Spreaders in the Twitter Spanish and English tasks at PAN@CLEF 2021. We present a hybrid system that uses Transformer-based models as feature extractors at the tweet level in combination with mixed pooling techniques. This approach allows computing the author's representative embeddings, which later fed an ML classifier. We explore the incorporation of features from Transformer-based models, Sentiment Analysis, and Hate lexicons to boost the feature extraction process. The results show that through this approach, it is possible to achieve 67% and 78% accuracy in the English and Spanish test datasets.

Keywords

Hate speech, Author profiling, Transformers, Mixed pooling

1. Introduction

Author profiling is part of digital text forensics and aims at determining the characteristics of the author of a document (i.e., age, gender, personality) [1]. As Online Social Networks (OSNs) grow, this task has become even more critical. As an example, platforms such as Twitter have recently experienced an increase in the use of abusive language and hate-based activities, partially promoted by the anonymity of its users, a fact that favours the presence of hate spreaders [2, 3]. The existence of online hate-spreaders undermines the integrity of online conversations by sharing inflammatory claims that influence public opinion and sow conflict on social or political issues [4, 2]. Therefore, the development of tools devoted to identifying hate-spreading at the author level is a new crucial challenge in the ever-evolving field of Artificial Intelligence.

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

alvaro.huertas.garcia@alumnos.upm.es (Á. Huertas-García); javier.huertas.tato@upm.es (J. Huertas-Tato); alejandro.martin@upm.es (A. Martín); david.camacho@upm.es (D. Camacho)

https://github.com/Huertas97 (Á. Huertas-García)

0000-0003-2165-0144 (Á. Huertas-García); 0000-0003-4127-5505 (J. Huertas-Tato); 0000-0002-0800-7632 (A. Martín); 0000-0002-5051-3475 (D. Camacho)

The scope of this paper is to describe a Natural Language Processing (NLP) based approach that makes use of Machine Learning (ML) and Deep Learning (DL) techniques for the PAN@CLEF21 Author Profiling Hate Speech Spreaders on Twitter Task [5, 6, 7]. This paper is organized into five sections. Section 2 provides a general view of some related works on author profiling and the description of the PAN 2021 task [5, 7]. Section 3 introduces our proposed approach. Section 4 describes the results from the experiments conducted. Finally, the conclusions are covered in Section 5.

2. Task Description and Related work

In recent years, there has been growing interest in author profiling and hate speech detection [3]. Since 2013, PAN organizers have proposed different tasks of author profiling in social media such as fake news spreader detection, bot detection, or age and gender characterization [8, 9]. The current task addressed in this paper of Author Profiling Hate Speech Spreaders on Twitter [5, 7] consists in determining whether an author spreads hate speech given its Twitter feed. The task adopts a multilingual perspective since the challenge includes both English and Spanish languages. For each language, the training data includes 200 authors with a Twitter feed of 200 tweets per author. For the English and Spanish tasks, the performance of the system is ranked in terms of accuracy as it is a binary classification, and the training data is balanced.

The complexity involved in natural language makes hate-speech detection a very challenging task [10] and requires a well-defined feature extraction process to infer the linguistic properties that enable hate-speech detection [3]. Regarding the feature extraction process, different studies have been dedicated to the use of Natural Language Processing (NLP) in combination with Machine Learning (ML) and Deep Learning (DL). Traditional feature extraction techniques such as Bag-of-words (BoW) [2], and ML algorithms, such as Support Vector Machines [11] and Naive Bayes [12], have been applied for hate-speech classification. As artificial intelligence techniques evolve, DL approaches were incorporated in this task, beating traditional state-of-the-art methods [10]. In [3] the authors propose a transfer learning approach for hate speech understanding using the unsupervised pretrained model BERT [13] fine-tuned for hate-speech. Moreover, Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks have also been proposed in the SemEval 2019 Task 5, also known as HatEval, focused on detecting hate speech against immigrants and women in Spanish and English at tweet-level [4]. Overall, multiple scientific projects have been dedicated to addressing the aspects of author profiling in social media with a focus on hate-speech detection, where NLP in combination with ML and DL methods have shown excellent results.

3. Profiling Hate Speech Spreaders with transformers and mixed pooling

This section overviews our approach for Hate Speech Spreaders profiling. During the feature extraction process, a Transformer-based model [14] is used in combination with a mixed pooling technique [15] that allows to extract a representative embedding for each user according to

his/her tweets. The embeddings are then used to train a Machine Learning classifier, which classifies the author as hate-spreader or non-hate-spreader.

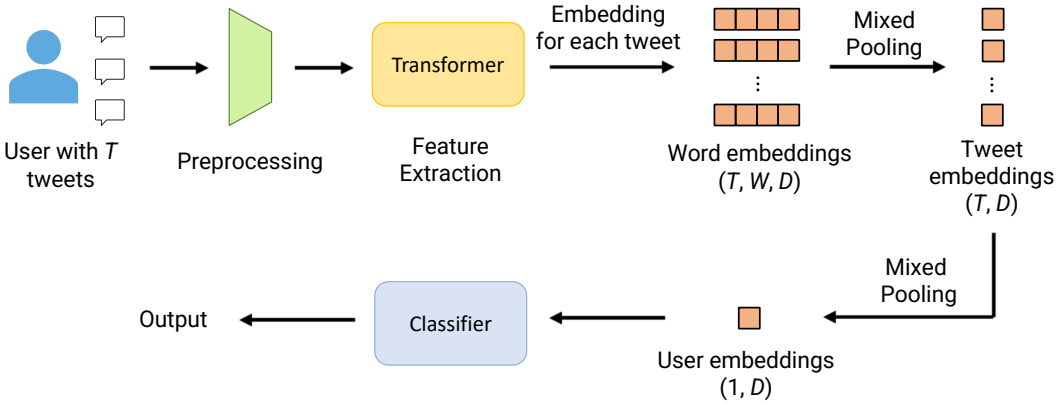

Figure 1: Diagram of the architecture of the proposed approach. T is the number of tweets per author; W is the number of words that each tweet is split into; and D is the number of model dimensions.

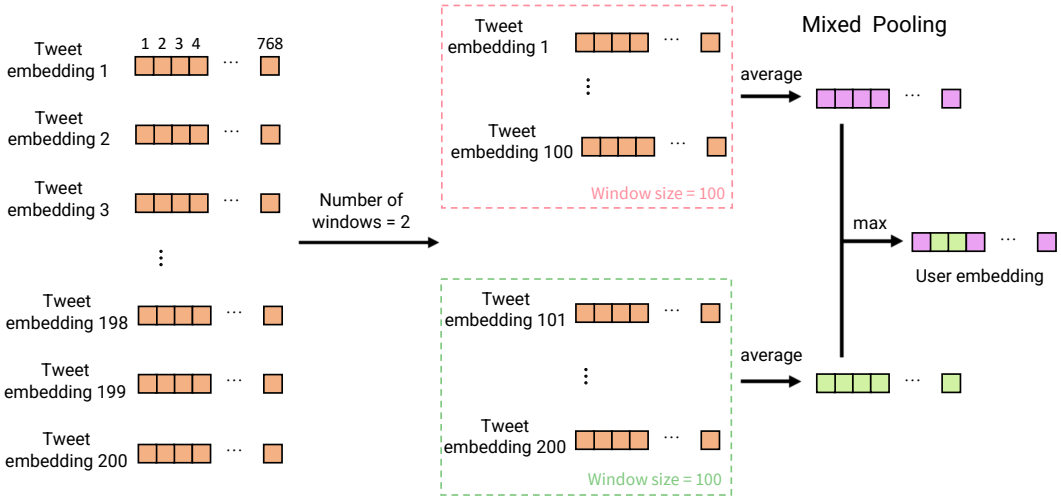

Figure 2: Example of mixed pooling process for condensing tweet embeddings into the user embedding.

3.1. Overview

A general overview of the proposed approach is provided in Figure 1. The first step is to preprocess the text data contained in each tweet. Then, we extract features from each preprocessed tweet using a pretrained transformer for each language. These models take as input a single tweet text and are fine-tuned to perform a binary classification where the input is labelled as hate or non-hate using the English or Spanish HatEval data [4], respectively. It is worth mentioning that the features used to represent each tweet are extracted from the output of the

last hidden layer before the classification layer. According to how the transformer architecture operates [13, 14], these features depict features for each word received as input. Thus, a pooling technique is required to condense these features into one representation, which can be then used to represent the whole tweet. A second mixed pooling is then applied to obtain a representation of the complete feed of each user, combining the representation of each of his/her tweets.

As previously mentioned, the pooling technique used for combining both the words representing one tweet and the whole author feed is mixed pooling. Yu et al. [15] introduced mixed pooling as a hybrid approach between average pooling and max pooling (see Figure 2). These authors proved the superiority of mixed pooling over max and average pooling techniques for image classification as it captures more local spatial information. As well as average and max pooling techniques are used in the NLP field, the mixed pooling technique can also infer information encoded in different embeddings into one single embedding [16, 17].

For the sake of simplicity, Figure 2 only describes the pooling process for condensing one feed of tweet embeddings into an user embedding. However, the same mixed pooling technique is also used for condensing word embeddings into tweet embedding. Firstly, the Twitter feed embeddings from a user are split into groups named windows. Secondly, average pooling is performed across each dimension of the window embedding. Finally, the user embedding is obtained by applying max pooling across the dimensions of the different average window embeddings computed in the previous step. Consequently, the number of windows is a hyperparameter that affects the outcome user embeddings that will be fed into a classifier, where mixed pooling is equal to max pooling when the window size (the number of items by window) is 1 and equal to average pooling when the window size is equal to the number of tweet feed (i.e., 200 tweets).

3.2. Pre-processing

Different preprocessing steps are applied for English and Spanish tasks. Although the PAN@CLEF21 task's data are already preprocessed at a certain level (URLs, Twitter's mentions, and hashtags are already normalized with special tokens), it is necessary to apply a more intense text preprocessing due to the noise introduced by Twitter slang. This preprocessing step is applied to both the PAN@CLEF21 data and the HatEval data used for training the transformer-based models.

English users' tweets are preprocessed using the *ftfy* package [18] to repair Unicode and emoji errors; *tweet-preprocessor* package[1] for deleting mentions, URLs, hashtags, and reserved characters (i.e., RT, FAV); and *ekphrasis* package [19] for normalizing percentages, time, dates, emails, phones and numbers. Contractions and emojis are not removed. For Spanish users' tweets, the same preprocessing steps are implemented adding a normalizing Spanish accent step.

3.3. Feature Extraction

Turning now to the transformer-based models used for feature extraction at tweet level, firstly, we established a baseline approach based on two models fine-tuned on English and Spanish

[1] https://github.com/s/preprocessor

HatEval tasks, respectively. Besides, feature enhancement was also tested by concatenating new features to the baseline from other resources.

3.3.1. Baseline approach

To select an English tweet-level feature extractor model based on transformers, we evaluated different models already fine-tuned and our own model *distilroberta-base* fine-tuned on the English HatEval task [4]. On the other hand, the already fine-tuned models were *bertweet-base-hate*, *twitter-roberta-base-hate*, *bertweet-base-offensive* and *twitter-roberta-base-offensive*, all of them belonging the to *cardiffnlp* group[2]. The transformer based models considered for Spanish feature extraction were *stsb-xlm-r-multilingual* and *distilbert-multilingual-nli-stsb-quora-ranking*, fine-tuned
and evaluated only with Spanish HatEval data, and *stsb-xlm-r-multilingual*, fine-tuned and evaluated with English and Spanish HatEval data. These multilingual models are part of the *sentence-transformers* group[3]. All these models are publicly available at Hugging Transformer API [20]. The model with the best performance in each language in terms of the official HatEval metric (macro-averaged F1-score) [4] is selected as the feature extractor.

3.3.2. Feature Enhancement approach

The feature enhancement approach appends new features to the baseline system by concatenating features from Vader-Sentiment-Analysis [21], Hatebase lexicon[4] and Detoxify [22]. The Vader-Sentiment-Analysis provide four new features: positive, negative and neutral scores, and an overall normalized score from the different sentiment lexicons ratings used by the Vader system named the compound score. It is only available for the English task. The Hatebase lexicon is a collection that contains multilingual parsed hateful lexicons used in the OSNs associated with a hate score. Therefore, Hatebase adds two new features, the frequency of hate lexicons detected in a user Twitter feed and the average hate score from these lexicons. Finally, a new feature is supplemented with the prediction of the Detoxify model [22]. Detoxify has a multilingual model version trained to predict the toxicity level of a comment on Jigsaw Multilingual Toxic Comment Classification challenge[5].

3.4. Author classification: Hyperparameter tuning and classifier

The classifiers tested were Naive Bayes (NB), Random Forest (RF), Logistic Regression with L1 and L2 regularization (LR1 and LR2, respectively), Elastic Net, and Support Vector Classifier (SVC) for both English and Spanish tasks, models that have shown excellent results in other researches [23].

To obtain the best results and avoid overfitting, we tuned the hyperparameters of the transformer-based models, the number of windows for mixed pooling, and the hyperparameters

[2]https://huggingface.co/cardiffnlp
[3]https://huggingface.co/sentence-transformers
[4]https://hatebase.org/
[5]https://kaggle.com/c/jigsaw-multilingual-toxic-comment-classification

Table 1
Hyperparameters search space used during the development of the proposed approach. C is the inverse of regularization strength while logspace is the logarithmic sequence (start base, end base, number of elements).

Optimization	Method	Hyperparameters	Values
Transformers	Grid and Bayesian search	learning rate	min = 1e-6 , max = 1e-3
		epochs	min = 1, max = 20
		weight decay	min = 0 , max = 1
		gradient accumulation steps	min = 1 , max = 4
		scheduler	linear schedule with warmup cosine schedule with warmup polynomial with warmup
Mixed Pooling	Grid search	number of windows	min = 1 , max = 200
RF	Grid search with CV = 5	n_estimators	[5, 10, 15, 30]
		max_depth	[3, 5, 10, 15, 20]
		min_samples_split	[2, 5, 10]
		min_samples_leaf	[1, 2, 4]
		max_features	[2, 3, "auto"]
		min_samples_split	[8, 10, 12]
LR1	Grid search with CV = 5	C	logspace(-3, 2, 8)
LR2	Grid search with CV = 5	C	logspace(-3, 2, 8)
Elastic Net	Grid search with CV = 5	C	logspace(-3, 2, 8)
		L1_ratio	[0 , 0.33333333, 0.66666667, 1]
SVC	Grid search with CV = 5	C	numpy logspace(-3, 2, 10)
		Kernel	polynomial, RBF, linear
		Gamma	logspace(-3, 3, 10)

of the classifiers using Grid and Bayesian search methods. The hyperparameters explored for each step of the proposed approach are summarized in Table 1.

4. Experiments and Results

This section presents the HatEval results used to select the feature extraction models, the number of windows hyperparameter optimization process, and the final approach presented for each language in the PAN@CLEF21 task.

4.1. Feature Extractor Model

Table 2 reports the performance on the English HatEval test set of transformer-based models evaluated as feature extractors. It can be seen that the already fine-tuned *bertweet-base-hate* from *cardiffnlp* group has the best values in terms of macro-average F1-score (70.39%) and Matthews correlation coefficient (41.38%). Bertweet [24] is a pre-trained language model for English tweets with the same architecture as BERT-base [13] trained using the RoBERTa pre-training procedure [25]. This model outperforms the topmost model from the original competition (65.10% F1-score) [4]. Moreover, this fine-tuned model scores the same as the complex feature enhancing approach proposed by Zhou et al. [26] composed of the concatenation of ELMo,

BERT base uncased and CNNs neural networks. Consequently, the *bertweet-base-hate* model is selected as the baseline method for English feature extraction at tweet level.

Regarding the Spanish HatEval test results summarized in Table 3, it can be seen that *stsb-xlm-r-mutilingual* model fine-tuned on hate speech with the Spanish HatEval training data has the best values, with 77.01% macro-average F1-score and 55.21% Matthews correlation coefficient. Reimers and Gurevych [17] developed this model using the teacher-student technique for distilling the knowledge from a monolingual model fine-tuned on STS Benchmark [27] into the XLM-RoBERTa model [28]. Remarkably, when we fine-tune this model on the Spanish HatEval task, it achieves better results than the topmost model from the original competition (73.00% F1-score). Therefore, this model is selected as the baseline method for Spanish feature extraction at tweet level.

Table 2
Performance of the transformer-based models evaluated as tweet-level feature extractor on the English HatEval test set. The (*) symbol represents our own fine-tuned models in this task. Performance is reported as macro-averaged F1- score and Matthews correlation coefficient (MCC) $\times$ 100.

Models	F1-macro	MCC
cardiffnlp/bertweet-base-hate	**70.39**	**41.38**
cardiffnlp/twitter-roberta-base-hate	69.40	39.25
mrm8488/distilroberta-finetunes-tweets-hate-speech*	59.39	32.17
cardiffnlp/bertweet-base-offensive	54.14	13.45
cardiffnlp/twitter-roberta-base-offensive	54.52	14.80

Table 3
Performance of the transformer-based models evaluated as tweet-level feature extractor on the HatEval task. Since the models are multilingual, the Language column indicates the languages used for their training and evaluation. Performance is reported as macro-averaged F1-score and Matthews correlation coefficient (MCC) $\times$ 100.

Models	Language	F1-macro	MCC
sentence-transformers/stsb-xlm-r-multilingual	**ES**	**77.01**	**55.21**
sentence-transformers/distilbert-multilingual-nli-stsb-quora-ranking	ES	69.55	44.64
sentence-transformers/stsb-xlm-r-multilingual	EN-ES	62.84	32.83

4.2. Window size hyperparameter tuning

Once the feature extraction models are selected, the mixed pooling number of windows hyperparameter is optimized. For that purpose, 25% of the PAN@CLEF21 training data were reserved as a development set in a stratified way (i.e., 50 of the 200 training authors).

Figure 3 and Figure 4 pinpoint the accuracy in the PAN@CLEF21 development set as a function of the number of windows. Before interpreting our results, we would like to restate that mixed pooling is equal to max-pooling when the number of windows is equal to the number of tweet embeddings (i.e., 200 tweets) and equal to the average pooling when the number of

windows is 1. Our results prove that, for both languages, mixed pooling is better than average and max pooling. The best performance is achieved with a number of windows equals to 26 with the SVC classifier in the English task (80.00%), and 32 with the LR1 classifier in the Spanish task (88.00%). Random Forest also achieves an 80% accuracy score for English. However, we opted for SVC because this type of classifier showed the best results in HatEval [4].

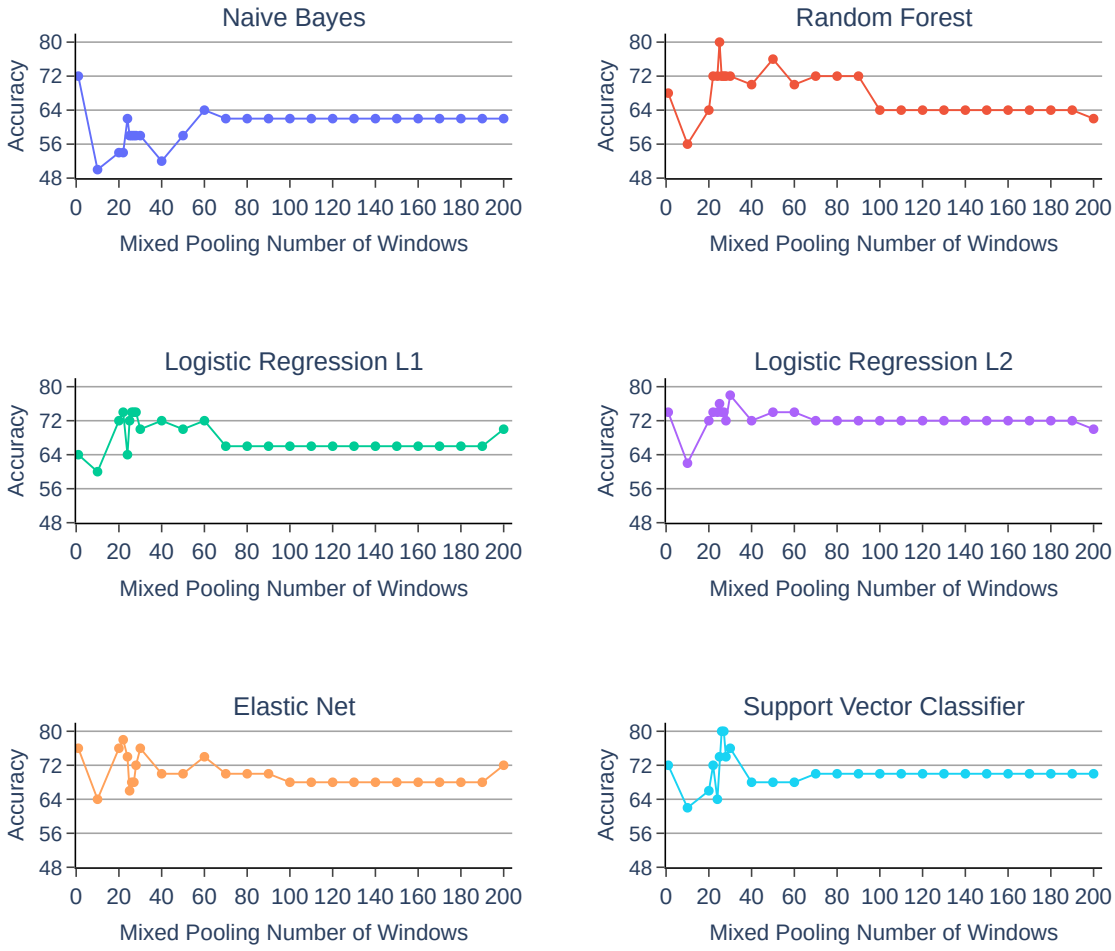

Figure 3: Hyperparameter optimization. Accuracy × 100 in PAN@CLEF21 English development set as a function of the number of windows for mixed pooling.

4.3. Baseline and Feature Enhancement approach Evaluation

In this section *EN-model-NW-26* and *ES-model-NW-32* will be used to refer to the English and Spanish baseline models with the window size hyperparameter selected based on the previous results.

The results of the author classification as hate spreaders on the PAN@CLEF21 development set are presented in Table 4 and Table 5. From the results obtained, it can be seen that the feature enhancement approach improves the *ES-model-NW-32* results when Detoxify model [22] is

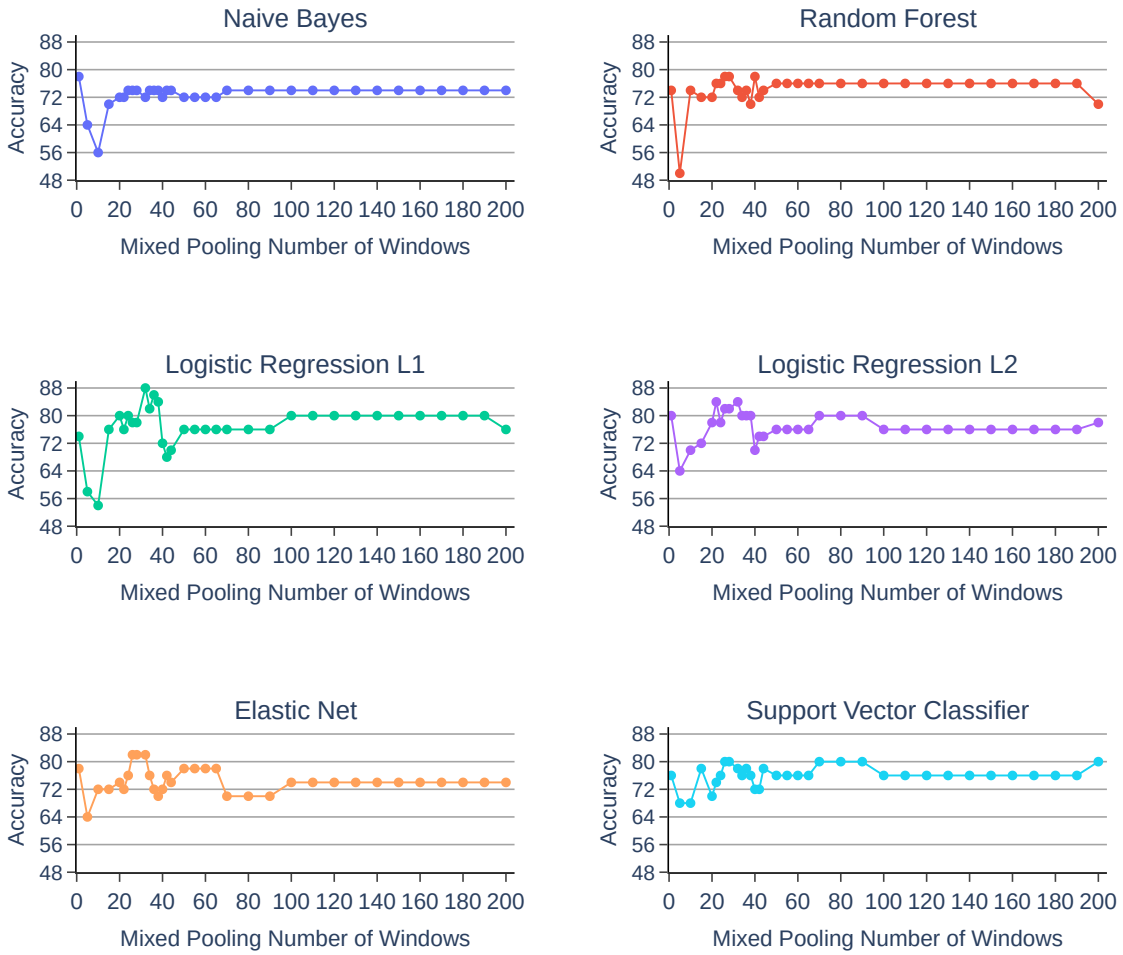

Figure 4: Hyperparameter optimization. Accuracy $\times$ 100 in PAN@CLEF21 Spanish development set as a function of the number of windows for mixed pooling.

included in combination with an LR1 classifier. On the other hand, *EN-model-NW-26* baseline approach is not improved with any of the features added.

Consequently, the feature enhancement approach composed of *stsb-xlm-r-mutilingual* fine-tuned on Spanish HatEval with 32 windows (*ES-model-NW-32*) and Detoxify as feature extractors, and a Logistic Regression with L1 regularization as hate spreaders classifier is the submitted approach for the Spanish task of PAN@CLEF21. Regarding the English task, the baseline approach composed of *bertweet-base-hate* with a number of windows equals to 26 (*EN-model-NW-26*) as feature extractor and a Support Vector Classifier as hate spreaders classifier is the one submitted. Concerning the classifiers, the Spanish LR1 classifier hyperparameter is $C = 0.139$, and the English SVC classifier hyperparameters values are $C = 0.001$ and $\gamma = 0.1$ with polynomial kernel.

Table 4
Performance of the proposed approach on the development set of the English PAN@CLEF21 Profiling Hate Speech Spreaders on Twitter Task. Performance is reported as Accuracy and Matthews correlation coefficient (MCC) $\times$ 100.

Feature Extractor	Classifier	Accuracy	MCC
EN-model-NW-26	SVC	**80.00**	**60.00**
EN-model-NW-26 + Hatebase	SVC	80.00	60.00
EN-model-NW-26 + Detoxify	SVC	80.00	60.00
EN-model-NW-26 + Vader	SVC	80.00	60.00
EN-model-NW-26 + Detoxify + Hatebase	SVC	80.00	60.00
EN-model-NW-26 + Detoxify + Vader	SVC	80.00	60.00
EN-model-NW-26 + Hatebase + Vader	SVC	80.00	60.00
EN-model-NW-26 + Detoxify + Hatebase + Vader	SVC	80.00	60.00

Table 5
Performance of the proposed approach on the development set of the Spanish PAN@CLEF21 Profiling Hate Speech Spreaders on Twitter Task. Performance is reported as Accuracy and Matthews correlation coefficient (MCC) $\times$ 100.

Feature Extractor	Classifier	Accuracy	MCC
ES-model-NW-32	LR1	88.00	76.99
ES-model-NW-32 + Hatebase	LR1	88.00	76.99
ES-model-NW-32 + Detoxify	LR1	**90.00**	**80.58**
ES-model-NW-32 + Detoxify + Hatebase	LR2	86.00	72.06

Table 6
Official results of the Profiling Hate Speech Spreaders on Twitter PAN@CLEF21 Task. Performance is reported as Accuracy $\times$ 100.

Feature Extractor	Classifier	Task	Accuracy
ES-model-NW-32 Detoxify	LR1	ES	**78.00**
EN-model-NW-26	SVC	EN	**67.00**

4.4. Official results

The official results in terms of accuracy for Profiling Hate Speech Spreaders on Twitter PAN@CLEF21 Task are reported in Table 6, reaching 78% accuracy in the Spanish task with the *ES-model-NW-32* + Detoxify as feature extractor and the Logistic Regression classifier with L1 regularization. In the English task, 67% accuracy is obtained with the *EN-model-NW-26* procedure as feature extractor and a Support Vector Machine classifier.

5. Conclusion

In this work, we proposed a Profiling Hater Spreader system for Twitter tasks in Spanish and English in PAN 2021. We presented a hybrid system composed of transformer-based models as tweet-level feature extractors, mixed pooling as pooling technique to compute author embeddings, and Machine Learning models as classifiers. Finally, we achieved an accuracy score of **78%** in Spanish and **67%** in English, leading to an average accuracy of **72.5%**. In future work, we will most likely test new pooling techniques, such as attentional pooling, and add more hate-labelled data to refine the feature extraction models and boost their performance.

Acknowledgements

This work has been partially supported by the following grants and funding agencies: Spanish Ministry of Science and Innovation under TIN2017-85727-C4-3-P (DeepBio) grant, by Comunidad Autónoma de Madrid under S2018/TCS-4566 grant (CYNAMON), and by BBVA FOUNDATION GRANTS FOR SCIENTIFIC RESEARCH TEAMS SARS-CoV-2 and COVID-19 under the grant: "*CIVIC: Intelligent characterisation of the veracity of the information related to COVID-19*". Relevant parts of this research is a result of the project IBERIFIER - Iberian Digital Media Research and Fact-Checking Hub, funded by the European Commission under the call CEF-TC-2020-2 (European Digital Media Observatory), grant number 2020-EU-IA-0252. Finally, the work has been supported by the Comunidad Autónoma de Madrid under Convenio Plurianual with the Universidad Politécnica de Madrid in the actuation line of "*Programa de Excelencia para el Profesorado Universitario*".

References

[1] M. Potthast, F. Rangel, M. Tschuggnall, E. Stamatatos, P. Rosso, B. Stein, Overview of pan'17, in: G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2017, pp. 275–290.

[2] P. Burnap, M. L. Williams, Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making: machine classification of cyber hate speech, Policy & Internet 7 (2015) 223–242. doi:10.1002/poi3.85.

[3] M. Mozafari, R. Farahbakhsh, N. Crespi, A bert-based transfer learning approach for hate speech detection in online social media, in: H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, L. M. Rocha (Eds.), Complex Networks and Their Applications VIII, Springer International Publishing, Cham, 2020, pp. 928–940.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. doi:10.18653/v1/S19-2007.

[5] 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.

[6] 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. doi:10.1007/978-3-030-22948-1_5.

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

[8] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Overview of pan 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2020, pp. 372–383.

[9] W. Daelemans, M. Kestemont, E. Manjavacas, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Tschuggnall, M. Wiegmann, E. Zangerle, Overview of pan 2019: Bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2019, pp. 402–416.

[10] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion (2017). doi:10.1145/3041021.3054223.

[11] S. Malmasi, M. Zampieri, Challenges in discriminating profanity from hate speech, 2018. arXiv:1803.05495.

[12] K. Demeyer, E. Lievens, J. Dumortier, Blocking and removing illegal child sexual content: analysis from a technical and legal perspective: blocking and removing illegal child sexual content, Policy & Internet 4 (2012) 1–23. doi:10.1002/poi3.8.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[15] D. Yu, H. Wang, P. Chen, Z. Wei, Mixed pooling for convolutional neural networks, in: D. Miao, W. Pedrycz, D. Ślęzak, G. Peters, Q. Hu, R. Wang (Eds.), Rough Sets and Knowledge Technology, Springer International Publishing, Cham, 2014, pp. 364–375.

[16] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. arXiv:1908.10084.

[17] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, 2020. arXiv:2004.09813.

[18] R. Speer, ftfy, Zenodo, 2019. doi:10.5281/zenodo.2591652, version 5.5.
[19] C. Baziotis, N. Pelekis, C. Doulkeridis, Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 747–754.
[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2020. arXiv:1910.03771.
[21] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: ICWSM, 2014.
[22] L. Hanu, Unitary team, Detoxify, Github. https://github.com/unitaryai/detoxify, 2020.
[23] J. Huertas-Tato, A. Martín, J. Fierrez, D. Camacho, Fusion of cnns and statistical indicators to improve image classification, arXiv preprint arXiv:2012.11049 (2020).
[24] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, 2020. arXiv:2005.10200.
[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.
[26] Y. Zhou, Y. Yang, H. Liu, X. Liu, N. Savage, Deep learning based fusion approach for hate speech detection, IEEE Access 8 (2020) 128923–128929. doi:10.1109/ACCESS.2020.3009244.
[27] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–14. doi:10.18653/v1/S17-2001.
[28] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.