Profiling Hate Speech Spreaders on Twitter Task at PAN 2021

Francisco Rangel[1], Gretel Liz De La Peña Sarracén[2], BERTa Chulvi[2], Elisabetta Fersini[3] and Paolo Rosso[2]

[1]*Symanto Research, Spain*
[2]*Universitat Politècnica de València, Spain*
[3]*Università Degli Studi di Milano-Bicocca, Italy*

Abstract
This overview presents the Author Profiling shared task at PAN 2021. The focus of this year's task is on determining whether or not the author of a Twitter feed is keen to spread hate speech. The main aim is to show the feasibility of automatically identifying potential hate speech spreaders on Twitter. For this purpose a corpus with Twitter data has been provided, covering the English and Spanish languages. Altogether, the approaches of 66 participants have been evaluated.

Keywords
hate speech, hate speech spreaders, author profiling, natural language processing, artificial intelligence

1. Introduction

Hate speech (HS) is commonly defined as any communication that disparages a person or a group on the basis of some characteristic, such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or others [1]. Given the huge amount of user-generated content on the Web and, in particular, on social media, the problem of detecting and, if possible, contrasting the HS diffusion, is becoming fundamental, for instance, in the fight against misogyny and xenophobia [2].

Having previously profiled bots [3] and fake news spreaders [4], at PAN'21 we have focused on profiling hate speech spreaders in social media, more specifically on Twitter, addressing the problem both in English and Spanish, as we did in the previous author profiling tasks. The goal has been to identify those Twitter users that can be considered haters, depending on the number of tweets with hateful content that they had spread. Our hypothesis is that users who do not spread hate speech may have a set of different characteristics compared to users who do. For example, they may use different linguistic patterns when they share posts compared to hate speech spreaders.

This is what we aim at investigating in this year's author profiling shared task where we address the problem of hate speech detection from the author profiling perspective. The final goal is profiling those authors who have shared some hate speech in the past. This will allow for identifying possible hate speech spreaders on Twitter as a first step towards preventing

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

CEUR Workshop Proceedings (CEUR-WS.org)

hate speech from being propagated among social media users. This should help for their early detection and, therefore, for preventing their further dissemination.

The remainder of this paper is organised as follows. Section 2 covers the state of the art, Section 3 describes the corpus and the evaluation measures, and Section 4 presents the approaches submitted by the participants. Sections 5 and 6 discuss results and draw conclusions respectively.

2. Related Work

Hate speech detection has received significant attention in the last few years. As a consequence, there are several available resources that can be used to identify hate speech. The survey in [5] presents a systematic analysis of a number of these resources. Corpora, lexicons and other resources are characterised in different dimensions. Moreover, there are other surveys that analyze different approaches used for hate speech detection [6, 7, 8, 9, 10, 11, 12]. In the vast amount of work, the approaches range from strategies based on linguistic features to more recent strategies based on deep learning. The work in [13] proposed a classifier based on a lexicon for hate speech detection. The lexicon is built with a rule-based approach that focuses on extracting subjective features. The experimental results suggested the inclusion of semantic and topic-based features. However, the use of machine learning approaches is recommended to improve the precision and recall scores. Moreover, other low-level features have been used, such as part-of-speech tags and dependency relationships [14, 13]. For example, the authors of [14] employed a supervised classification method that considers different features that can be divided into four classes: n-grams, linguistic, syntactic and distributional semantics.

On the other hand, several systems use traditional machine learning models such as support vector machines (SVM) and logistic regression (LR) [15, 11, 16]. The authors of [15] proposed an approach based on SVM that achieved state-of-the-art performance. It was simple and produced decisions easier to interpret than neural network methods according to the analysis of the authors. Other systems are based on deep learning models such as convolutional neural network (CNN) and long short term memory (LSTM) networks, including also attention mechanisms [17, 18, 19, 20, 10, 21, 22]. Moreover, some works proposed ensembles of neural networks. For example, the authors in [23] combined ten CNNs with different initialisation weights in an ensemble to improve the detection of hate speech. Among the deep learning models, the bidirectional encoder representations from Transformers (BERT) has been widely used [24, 25, 16, 26, 27]. The work presented in [28] investigates the potentialities of BERT for hate speech detection within social media by transfer learning with fine-tuned methods. The results showed that this strategy obtains a considerable performance in comparison to other existing approaches.

While most of the approaches focus on detecting whether a text is hateful or not, few works focus on the user account level detection. In [29], the authors studied the flow of posts generated by users on Gab[1]. The authors analysed the profiles and network of hateful and non-hateful users, focusing on the diffusion dynamics of hateful users. The observations suggested that hateful content propagates farther, wider and faster. Unlike this work, where the analysis was

[1]https://gab.com

carried out statically, the authors of [30] based their research on graphs but focusing on dynamic graphs and investigated the temporal effects of hate speech. The authors of [31] presented a comparative study of hate speech users on Twitter. The authors investigated the distinctive characteristics of hateful users and targeted users in terms of their profile, activities, and online visibility. They found that hateful users can be more popular and that participating in hate speech can result in greater online visibility.

The authors of [32] also focused on profiling hate speech spreaders on Twitter. This study used a methodology to obtain a graph given the entire profile of users, and investigated the difference between HS and non-HS spreaders in terms of activity patterns, word usage and network structure. The authors observed that hateful users are densely connected, thus they focused on exploiting the network of connections. The authors of [33] proposed a model that considers intra-user and inter-user representation learning for hate speech detection. The authors of [34] studied the use of emojis in white nationalist conversation on Twitter and observed the difference between the 'pro' and 'anti' stance.

During the last three years, at PAN we focused on profiling users who spread harmful information, as well as profiling bots due to their key role in its propagation on Twitter. Concretely, in 2019 the goal was discrimination bots from humans [35], in 2020 identifying possible fake news spreaders [36], and this year profiling haters.

Several have been the approaches that have been developed in these author profiling tasks. In 2019, 59 teams participated in the author profiling shared task and most of them used traditional approaches, such as SVM and logistic regression. Few teams proposed deep learning methods, such as convolutional neural network (CNN) and recurrent neural network (RNN). In [37], the authors combined both architectures based on character and word n-gram models and the authors pointed out the superiority of the model over individual classifiers. The authors in [38] used feedforward neural networks after an analysis of different methods, including a 2D CNN fed with large kernels matching with the number of features extracted from the TFIDF representation. The authors of [39] used the BERT model but it did not obtain a good position in the final ranking of the task.

In 2020, 66 teams participated in the author profiling shared task, and similarly to the previous year, most of the proposals were based on SVM and deep learning models like CNN and RNN. The authors of [40] and [41] obtained the best results for English and Spanish respectively. The authors of [40] employed a LR-based approach and the authors of [41] a SVM-based one.

3. Evaluation Framework

The purpose of this section is to introduce the technical background. We outline the construction of the corpus, introduce the performance measures and baselines, and describe the software submissions.

3.1. Corpus

To build the PAN-AP-2021 corpus[2] we have proceeded as follows. Firstly, we have looked for users considered potential haters. To do so, we have followed two approaches: *(i)* a keyword-based one (e.g. searching for hateful words mainly towards women or immigrants); and *(ii)* a user-based one, by inspecting users known as haters (e.g. users appearing in reports and/or press) and following their networks (followers and followees). Secondly, for the identified users, we have collected their timelines and manually annotated those tweets conveying hate. Thirdly, we have labelled as "keen to spread hate speech" those users with more than ten hateful tweets. Finally, we have collected two hundred tweets per Twitter user to build up the final dataset.

Table 1
Number of authors in the PAN-AP-21 corpus created for this task.

Language	Training	Test	Total
English	200	100	300
Spanish	200	100	300

Table 1 presents the statistics of the corpus that consists of 300 authors for each of the two languages, English and Spanish. For each author, we retrieved via the Twitter API her last 200 tweets. The corpus for each language is completely balanced, with 150 authors for each class (hate and non-hate speech spreaders), as well as by the number of tweets per user. We have split the corpus into training and test sets, following a proportion of 2/3 for training and the rest for test.

3.2. Performance Measure

The performance of the systems has been ranked by accuracy. For each language, we calculated individual accuracy for discriminating between the two classes. Finally, we averaged the accuracy values per language to obtain the final ranking.

3.3. Baselines

As baselines to compare the performance of the participants with, we have selected:

- Character n-grams with n ranging from 2 to 6 and Logistic Regression;
- Word n-grams with n ranging from 1 to 3 and SVM;
- Universal Sentence Encoder (USE) to feed a BiLSTM with;
- XLM-Roberta (XLMR) transformer to feed a BiLSTM with;
- Multilingual BERT (MBERT) transformer to feed a BiLSTM with;
- TFIDF vectors representing each user's text to feed a BiLSTM with.

[2]We should highlight that we are aware of the legal and ethical issues related to collecting, analysing and profiling social media data [42] and that we are committed to legal and ethical compliance in our scientific research and its outcomes. Concretely, we have anonymised the user name, masked all the user mentions and we did not use explicit labels for HS and non-HS spreaders.

- *LDSE* [43]. This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDSE is a weight, representing the probability of a term to belong to one of the different categories: HS vs. non-HS spreader. The distribution of weights for a given document should be closer to the weights of its corresponding category.

3.4. Software Submissions

Similar to previous year[3], we asked for software submissions. Within software submissions, participants submit executables of their author profiling software instead of just the output of their software on a given test set. For the software submissions, the TIRA experimentation platform was employed [44, 45], which renders the handling of software submissions at scale as simple as handling run submissions. Using TIRA, participants deploy their software on virtual machines at our site, which allows us to keep them in a running state [46].

4. Overview of the Submitted Approaches

This year, 66 teams participated in the author profiling shared task and 32 of them submitted the notebook paper. We analyse their approaches from three perspectives: preprocessing, features used to represent the authors' texts and classification approaches.

4.1. Preprocessing

With the aim at preventing bias towards some URLs, user mentions or hashtags, the corpus was provided with these elements already masked. In the same vein, some participants cleaned other Twitter-specific elements such as RT, VIA, and FAV[4] reserved words [47, 48, 49, 50, 51, 52], as well as emojis and other non-alphanumeric characters [53, 53, 50, 54, 51, 52, 55, 56, 57, 58], numbers [47, 48, 51, 52] or punctuation signs [57, 59, 60, 56, 51, 53, 47, 58, 50]. Various participants lower-cased the texts [53, 53, 47, 58, 54, 52, 55, 60], removed stop words [53, 58, 50, 52] or treated character flooding [55, 52, 47, 53]. Finally, some users got rid of short texts [52], stemmed or lemmatised [52, 53, 58, 50] and tokenised [58, 50] the tweets. Alcañiz and Andrés [53] expanded contractions and colloquial tokens (e.g., in Spanish, "x" was expanded to "por", and "k" to "que"), as well as Huertas-García et al. [48] who also considered slang words. Some participants carried out statistical analysis with preprocessing purposes. For instance, t-SNE has been used by Finogeev et al. [61] and Ceron and Casula [62]. These last authors also applied the Kolmogorov-Smirnov test. Finally, Ikae [63] used TFIDF and shift graphs [64].

4.2. Text Representation

The participants have used a high variety of different features and their combinations, albeit we can group them into the following main groups: *(i)* n-grams; *(ii)* stylistics; *(iii)* personality and

[3]Due to some technical issues due to TIRA power outage, this year we have also allowed some users to directly sent us their prediction files as well as their software for us to reproduce their systems.

[4]RT is the acronym for *retweet*; VIA is a way to give the authorship to a user (e.g., "via @kicorangel"); and FAV stands for *favourite*.

emotions; and *iv)* deep learning-based such as embeddings.

As every year, one of the most used features has been the combination of n-grams (character- and word-based, but also syntactical ones, mostly weighted by TFIDF) [65, 58, 66, 63, 67, 56, 53]. In some cases, the participants also combined different types of n-grams with stylistic features and dictionary-based ones [68] or emotional- and personality-based ones. In this sense, Cervero [69] combined n-grams with big five personality traits [70], LIWC[5], lexicon-based emotions (NRC [71]), besides stylistic features.

Emotional-based features have been used by different participants, usually combined with other representations such as embeddings [59] and transformers [48]. In this regard, some participants have approached the task with specific and specialised lexicons related to hate speech. For instance, Lai et al. [72] have combined hate speech detection, user morality, named entities and communicative behaviour. Similarly, Tosev and Gievska [73] have combined hateful lexicon-related features with n-grams, sentiment and typed dependencies.

Deep learning-based representations such as embeddings or transformers have been widely used by the participants. For instance, word embeddings have been used by the authors of [74, 53, 47], in some cases also combined with stylistic features [74]. In other cases, the participants have used other kinds of embeddings, such as in the case of Cabrera et al. [57] who used semantic-emotion-based embeddings, or Puertas and Martínez-Santos [59] who combined lexical, statistical, syntactical, and phonetical-based embeddings.

Nevertheless, most of the participants used the latest transformer-based approaches. BERT has been the most widely used transformer [48, 61, 75, 76, 54, 62, 77, 60], in some cases at sentence level (SBERT) [49, 52], but also in some of its variations: ALBERT [75], RoBERTa [76, 54, 55], BERTTweet [54] and BETO [54].

Finally, it is worth mentioning the following authors for having approached the task from different perspectives. Babaei et al. [51] have adapted LDSE to work at character level. Labadie et al. [78] fine-tuned a transformer to modify the Impostor Method previously used in Authorship Verification. Irani et al. [79] aggregated document topics and combined them with ELMo representations to represent the users. Höllig et al. [80] used a CNN to extract features from external data they collected for this purpose. Finally, Puertas and Martínez-Santos [59] used, among others, phonetic embeddings to represent the Twitter users.

4.3. Machine Learning Classifiers

Regarding the classification approaches, most participants used traditional machine learning methods such as SVM [53, 47, 65, 61, 80, 58, 51, 62, 52, 55, 56, 57], Random Forest [79, 59, 65], LR [61, 55, 60], Adaboost [59], and a Ridge Classifier [80]. Some participants have evaluated several approaches together, such as Anwar [54] who used AutoML, Jain et al. [50] who tested Naive Bayes, KNN, SVM, Logistic Regresion, as well as deep learning methogs such as LSTM and BERT, or Katona et al. [68] who evaluated Logistic Regression, Random Forest, SVM and XGBoost. Many participants this year approached the task with ensembles of classifiers. For instance, Huertas-García et al. [48] combined SVC, Naive Bayes, Random Forest, Logistic Regression and Elastic Net. Cervero [69] ensembled Logistic Regression, SVC and Random Forest.

[5] http://www.liwc.net

Gómez and Hinojosa combined SVC, Random Forest and Logistic Regression, and Ikae [63] used Linear Discriminant Analysis, Gradient Boosting, Gaussian Naive Bayes, Bernoulli Naive Bayes, Random Forest and AdaBoost. Balouchzahi et al. [67] combined SVM, Logistic Regression and Random Forest while Tosev and Gievska [73] combined SVM, Logistic Regression and Random Forest.

Regarding deep learning approaches, Martín et al. [47] describe a custom architecture based on several layers as well as Baris and Magnossao [49] added several linear layers after BERT. Pallares and Herrero [81] used a Recurrent Neural Network while Siino et al. [66] used a Convolutional Neural Network. LSTM has been used by Uzan and HacoHen-Kerner [77] and BiLSTM by Vogel and Meghana [52].

5. Evaluation and Discussion of the Results

In this section, we present the results of the shared task, as well the analysis of the most common errors made by the best performing teams. Although we recommended to participate in both languages (English and Spanish), some participants only participated in English. We present the results for the two languages, finally showing the ranking computed according to the average between them.

5.1. Global Ranking

In Table 2, the overall performance of the participants is presented. The results are shown in terms of accuracy for both languages, and the ranking is its average.

The top-ranked participants approached the task as follows. The overall best result (79%) has been obtained by [66] with a 100-dimension word embedding representation to feed a CNN. The two ex aequo second best performing teams, respectively fine-tuned a transformer to replicate and modify the Impostor Method previously used in Authorship Verification [78], and used a meta-classifier fed with combinations of n-grams [67]. The fourth performing team [54] experimented with several transformer-based approaches such as BERT, BERTTweet, RoBERTa, BETO, and evaluated several machine learning approaches with AutoML. The seventh best performing team combined BERT with SVM. The following ones used SVM with features extracted with CNN [80], BERT and SVM [61], TFIDF and SVM [53], an ensemble of classifiers with a combination of n-grams, big five, LIWC, NRC lexicon and stylistic features [69], and a combination of n-grams, stylistic and dictionary-based features with a combination of Logistic Regression, Random Forest, SVM, and XgBoost classifiers [68].

The best results have been obtained in Spanish (85% vs. 75%). The best result in Spanish (85%) has been also obtained by [66], while the best result in English (75%) has been obtained by [60] with BERT and Logistic Regression.

As can be seen in Figure 1 and Table 3, the results for Spanish are higher than for English both in terms of average (77.98% vs. 70.66%) and maximum (85% vs. 79%) accuracies. Although the standard deviation is similar for both languages (5.39% vs. 5.25%), the inter-quartile range is larger for English (6% vs. 4%), showing a slightly more sparse distribution in this last language. This might be due to the highest number of outliers in the Spanish distribution, as shown in Figure 2.

Table 2
Overall accuracy of the submission to the task on profiling hate speech spreaders on Twitter: teams that participated in both languages (English and Spanish) are ranked by the average accuracy between both languages, teams that participated only in English (bottom right) are ranked by the accuracy on English. The best results for each language are highlighted in bold.

	PARTICIPANT	EN	ES	AVG
1	SiinoDiNuovo [66]	0.730	**0.850**	**0.790**
2	MUCIC [67]	0.730	0.830	0.780
2	UO-UPV [78]	0.740	0.820	0.780
4	andujar	0.720	0.820	0.770
4	anitei	0.720	0.820	0.770
4	anwar [54]	0.720	0.820	0.770
7	pagnan [62]	0.730	0.800	0.765
	LDSE [20]	0.700	0.820	0.760
	char nGrams+LR	0.690	0.830	0.760
8	hoellig [80]	0.730	0.790	0.760
9	bañuls	0.680	0.830	0.755
9	supaca	0.690	0.820	0.755
9	oleg [61]	0.670	0.830	0.750
9	moreno [53]	0.690	0.810	0.750
9	cervero [69]	0.700	0.800	0.750
14	katona [68]	0.700	0.790	0.745
	word nGrams+SVM	0.650	0.830	0.740
15	bagdon [55]	0.670	0.810	0.740
15	das [58]	0.670	0.810	0.740
17	ikae [63]	0.660	0.810	0.735
17	mata	0.700	0.770	0.735
19	lai [72]	0.620	0.840	0.730
19	jain [50]	0.660	0.800	0.730
19	villarroya	0.670	0.790	0.730
19	mktung	0.640	0.820	0.730
19	sercopa	0.670	0.790	0.730
19	castro	0.670	0.790	0.730
25	giglou [51]	0.650	0.800	0.725
25	huertas [48]	0.670	0.780	0.725
25	wentao	0.680	0.770	0.725
28	rus	0.610	0.830	0.720
28	tudo	0.650	0.790	0.720
30	jaiferhu	0.610	0.820	0.715
30	joshi	0.650	0.780	0.715
32	valiense [65]	0.630	0.790	0.710
32	krstev	0.650	0.770	0.710
34	martin [47]	0.650	0.770	0.710
35	gomez [74]	0.580	0.830	0.705
35	bakhteev	0.580	0.830	0.705

	Participant	En	Es	Avg
35	MaNa	0.640	0.770	0.705
38	cabrera [57]	0.620	0.780	0.700
38	esam [76]	0.630	0.770	0.700
38	zhang	0.630	0.770	0.700
41	dudko	0.610	0.780	0.695
41	meghana [52]	0.640	0.750	0.695
43	rubio	0.590	0.790	0.690
43	uzan [77]	0.620	0.760	0.690
45	herrero [81]	0.570	0.800	0.685
46	puertas [59]	0.600	0.760	0.680
	USE-LSTM	0.560	0.790	0.675
	XLMR-LSTM	0.620	0.730	0.675
47	ipek [49]	0.580	0.770	0.675
47	schlicht21	0.580	0.770	0.675
47	peirano	0.590	0.760	0.675
47	russo	0.550	0.800	0.675
	MBERT-LSTM	0.590	0.750	0.670
51	kazzaz	0.550	0.770	0.660
52	dorado	0.600	0.710	0.655
53	kobby [75]	0.530	0.770	0.650
53	kern	0.540	0.760	0.650
53	espinosa [56]	0.640	0.660	0.650
56	labadie	0.510	0.780	0.645
57	silva	0.560	0.690	0.625
57	garibo	0.570	0.680	0.625
59	estepicursor	0.510	0.720	0.615
60	spears	0.520	0.680	0.600
	TFIDF-LSTM	0.610	0.510	0.560
61	barbas	0.460	0.500	0.480

	Participant	En
62	dukic [60]	**0.750**
63	tosev [73]	0.700
64	amir [79]	0.680
65	siebert	0.680
66	iteam	0.650

Table 3
Statistics on the accuracy per language.

STAT	EN	ES	AVG
Min	0.4600	0.5000	0.4800
Q1	0.5925	0.7700	0.6800
Median	0.6500	0.7900	0.7150
Mean	0.6377	0.7798	0.7066
SDev	0.0643	0.0539	0.0524
Q3	0.6800	0.8100	0.7400
Max	0.7500	0.8500	0.7900
Skewness	-0.4719	-2.6820	-1.4672
Kurtosis	2.7391	13.4989	7.1170
Normality (p-value)	0.2264	1.129e-08	0.0148

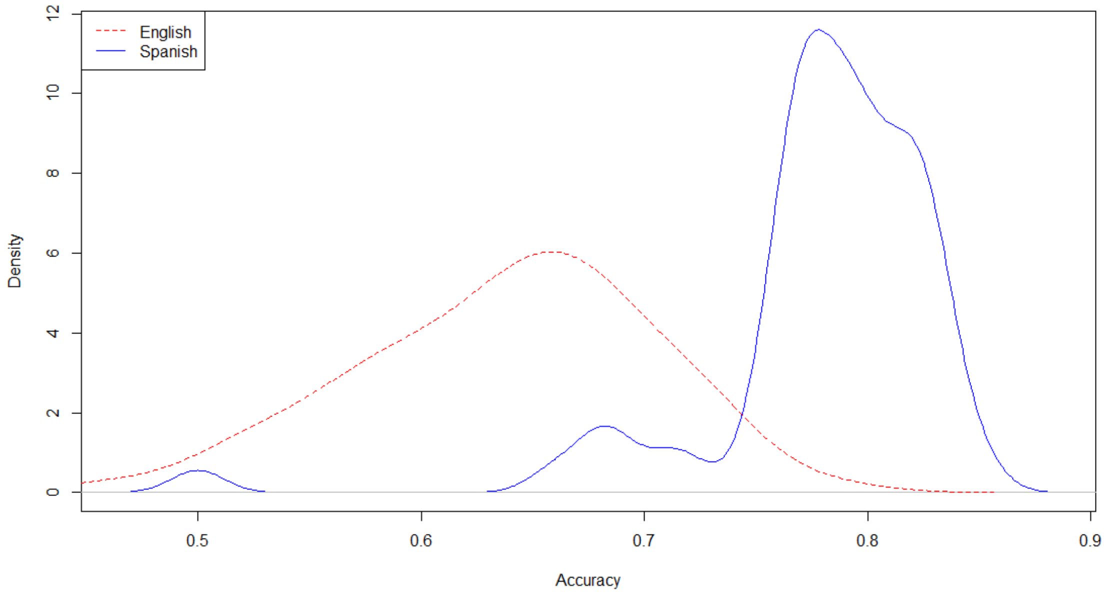

Figure 1: Density of the results in the different languages.

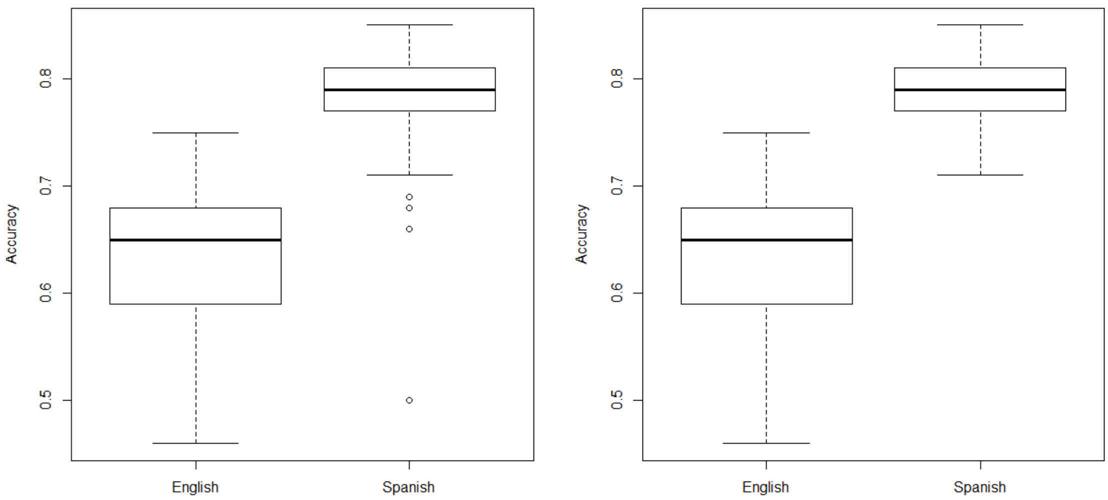

Figure 2: Distribution of results in the different languages. The figure on the left represents all the systems. The figure on the right removes the outliers.

5.2. Error Analysis

We have aggregated all the participants' predictions for HS vs. non-HS spreaders discrimination task, except baselines, and plotted the respective confusion matrices for English and Spanish in Figures 3 and 4, respectively. Figure 3 shows the confusion matrix for English. It can be seen that the error is highly balanced with respect to both classes, although slightly higher in the case of false positives (from non-HS to HS spreaders): 39.23% vs. 34.20%. Figure 4 shows the confusion matrix for Spanish, where the difference between false positives and false negatives is much higher (almost double): 28.13% vs. 16.85%. This higher number of false positives, mainly in the case of Spanish, is something to be investigated further in future research.

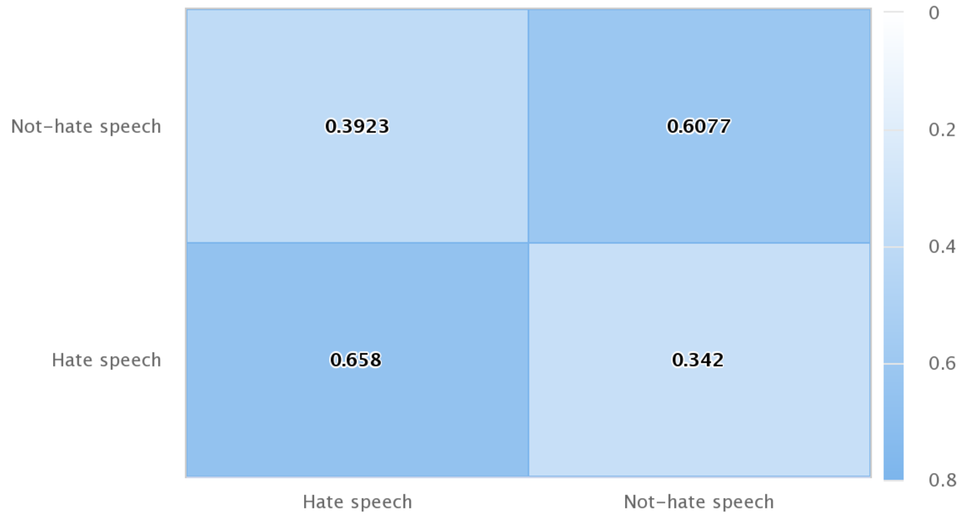

Figure 3: Aggregated confusion matrix for HS vs. non-HS spreaders in English.

5.3. Best Results

In Table 4 we summarise the best results per language. The best result in English (0.75) has been obtained with BERT and Logistic Regression. The best result in Spanish (0.85) has been obtained with a 100-dimension word embedding representation to feed a Convolutional Neural Network.

Table 4
Best results per language.

English	Spanish
Dukić and Sović [60] (0.75)	Siino et al. [66] (0.85)

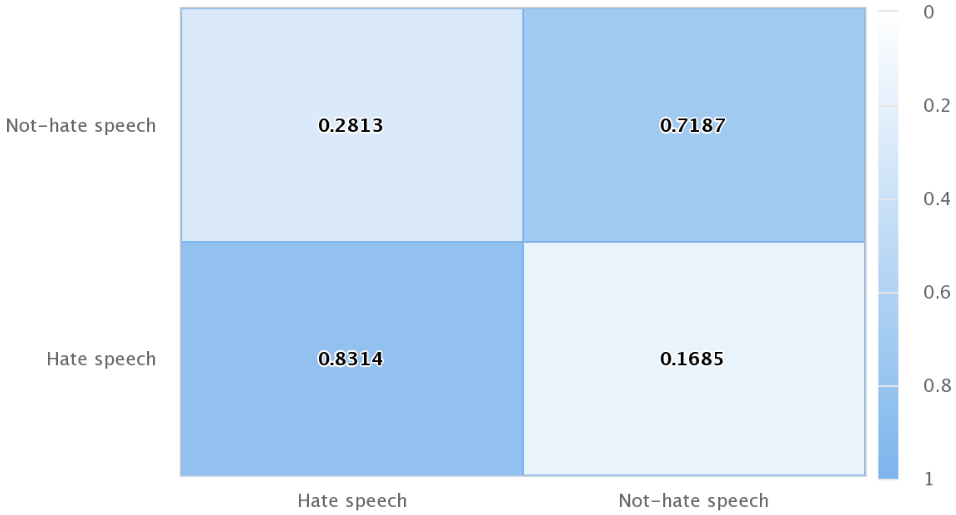

Figure 4: Aggregated confusion matrix for HS vs. non-HS spreaders in Spanish.

6. Conclusion

In this paper, we have presented the results of the 9th International Author Profiling Shared Task at PAN 2021, hosted at CLEF 2021. The participants had to discriminate on Twitter between HS and non-HS spreaders. The provided data cover the English and Spanish languages.

The participants used different features to address the task, mainly: *(i)* n-grams; *(ii)* stylistics; *(iii)* personality and emotions; and *(iv)* deep learning-based representations such as embeddings and mainly transformers. Concerning machine learning algorithms, the most used ones were Support Vector Machines and Logistic Regression, as well as ensembles of several algorithms. Nevertheless, few participants approached the task with only deep learning techniques. In such cases, they used Fully-Connected Neural Networks, RNN, CNN, LSTM and Bi-LSTM. According to the results and unlike previous years, the combination of traditional machine learning methods such as SVM or Logistic Regression with novel deep learning-based representations such as BERT and its variations have obtained the highest results.

The best results have been obtained in Spanish (0.85) by Siino et al. [66] with a 100-dimension word embedding representation to feed a Convolutional Neural Network. The best result in English (0.75) has been obtained by Dukić and Sović [60] with BERT and Logistic Regression. The overall best result (0.79) has been obtained by the first team.

The error analysis shows that the highest confusion in both languages is towards Hate Speech Spreaders (false positives), although the proportion with respect to false negatives in Spanish (28.13% vs. 16.85%) is much higher than in English (39.23% vs. 34.20%).

Looking at the results and the error analysis, we can conclude that: *(i)* it is feasible to automatically identify potential Hate Speech Spreaders on Twitter with high precision, even when only textual features are used; but *(ii)* we have to bear in mind *false positives* since especially in Spanish, they are almost double than false negatives, and misclassification might lead to ethical or legal implications [42].

Acknowledgments

First of all, we thank the participants: again 66 this year, as the previous year on Profiling Fake News Spreaders! We have to thank also Martin Potthast, Matti Wiegmann, Nikolay Kolyada, and Magdalena Anna Wolska for their technical support with the TIRA platform. We thank Symanto for sponsoring again the award for the best performing system at the author profiling shared task. The work of Francisco Rangel was partially funded by the Centre for the Development of Industrial Technology (CDTI) of the Spanish Ministry of Science and Innovation under the research project IDI-20210776 on Proactive Profiling of Hate Speech Spreaders - PROHATER (Perfilador Proactivo de Difusores de Mensajes de Odio). The work of the researchers from Universitat Politècnica de València was partially funded by the Spanish MICINN under the project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31), and by the Generalitat Valenciana under the project DeepPattern (PROMETEO/2019/121). This article is also based upon work from the Dig-ForAsp COST Action 17124 on Digital Forensics: evidence analysis via intelligent systems and practices, supported by European Cooperation in Science and Technology.

References

[1] J. T. Nockleby, Hate speech, in: Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan), pp. 1277–1279, 2000.

[2] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. Rangel, P. Rosso, M. Sanguinetti, SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, in: Proc. of the 13th Int. Workshop on Semantic Evaluation (SemEval-2019), co-located with the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), 2019.

[3] F. Rangel, P. Rosso, Overview of the 7th author profiling task at pan 2019: Bots and gender profiling., in: CLEF 2019 Labs and Workshops, Notebook Papers, 2019.

[4] F. Rangel, A. Giachanou, B. Ghanem, P. Rosso, Overview of the 8th Author Profiling Task at PAN 2019: Profiling Fake News Spreaders on Twitter, in: CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, 2020.

[5] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review, Language Resources and Evaluation (2020) 1–47.

[6] 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.
[7] A. Al-Hassan, H. Al-Dossari, Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus, in: 6th International Conference on Computer Science and Information Technology, 2019.
[8] U. Naseem, I. Razzak, P. W. Eklund, A Survey of Pre-processing Techniques to Improve Short-Text Quality: A Case Study on Hate Speech Detection on Twitter, Multimedia Tools and Applications (2020) 1–28.
[9] R. Cao, R. K.-W. Lee, T.-A. Hoang, DeepHate: Hate Speech Detection Via Multi-Faceted Text Representations, in: 12th ACM Conference on Web Science, 2020, pp. 11–20.
[10] A. Magalhaes, Automating Online Hate Speech Detection: A Survey of Deep Learning Approaches, Master's thesis, School of Informatics, University of Edinburgh, 2019.
[11] P. Fortuna, S. Nunes, A Survey on Automatic Detection of Hate Speech in Text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.
[12] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, arXiv preprint arXiv:1703.04009 (2017).
[13] N. D. Gitari, Z. Zuping, H. Damien, J. Long, A Lexicon-based Approach for Hate Speech Detection, International Journal of Multimedia and Ubiquitous Engineering 10 (2015) 215–230.
[14] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive Language Detection in Online User Content, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.
[15] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate Speech Detection: Challenges and Solutions, PloS one 14 (2019) e0221152.
[16] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020), arXiv preprint arXiv:2006.07235 (2020).
[17] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep Learning for Hate Speech Detection in Tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759–760.
[18] B. Gambäck, U. K. Sikdar, Using Convolutional Neural Networks to Classify Hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.
[19] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, N. Asokan, All You Need is" Love" Evading Hate Speech Detection, in: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, 2018, pp. 2–12.
[20] J. H. Park, P. Fung, One-step and Two-step Classification for Abusive Language Detection on Twitter, arXiv preprint arXiv:1706.01206 (2017).
[21] Z. Zhang, D. Robinson, J. Tepper, Detecting Hate Speech on Twitter using a Convolution-GRU based Deep Neural Network, in: European semantic web conference, Springer, 2018, pp. 745–760.
[22] G. L. De la Pena Sarracén, R. G. Pons, C. E. M. Cuza, P. Rosso, Hate Speech Detection using Attention-based LSTM, EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018) 235.
[23] S. Zimmerman, U. Kruschwitz, C. Fox, Improving Hate Speech Detection with Deep

Learning Ensembles, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[24] 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.
[25] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the HASOC Track at Fire 2019: Hate Speech and Offensive Content Identification in Indo-European Languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 14–17.
[26] M. Mozafari, R. Farahbakhsh, N. Crespi, Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model, PloS one 15 (2020) e0237861.
[27] N. S. Samghabadi, P. Patwa, P. Srinivas, P. Mukherjee, A. Das, T. Solorio, Aggression and Misogyny Detection Using BERT: A Multi-task Approach, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 2020, pp. 126–131.
[28] M. Mozafari, R. Farahbakhsh, N. Crespi, A BERT-based Transfer Learning Approach for Hate Speech Detection in Online Social Media, in: International Conference on Complex Networks and Their Applications, Springer, 2019, pp. 928–940.
[29] B. Mathew, R. Dutt, P. Goyal, A. Mukherjee, Spread of Hate Speech in Online Social Media, in: Proceedings of the 10th ACM conference on web science, 2019, pp. 173–182.
[30] B. Mathew, A. Illendula, P. Saha, S. Sarkar, P. Goyal, A. Mukherjee, Hate Begets Hate: A Temporal Study of Hate Speech, Proceedings of the ACM on Human-Computer Interaction 4 (2020) 1–24.
[31] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, E. Belding, Peer to Peer Hate: Hate Speech Instigators and Their Targets, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 12, 2018.
[32] M. Ribeiro, P. Calais, Y. Santos, V. Almeida, W. Meira Jr, Characterizing and Detecting Hateful Users on Twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 12, 2018.
[33] J. Qian, M. ElSherief, E. M. Belding, W. Y. Wang, Leveraging Intra-user and Inter-user Representation Learning for Automated Hate Speech Detection, arXiv preprint arXiv:1804.03124 (2018).
[34] L. Hagen, M. Falling, O. Lisnichenko, A. A. Elmadany, P. Mehta, M. Abdul-Mageed, J. Costakis, T. E. Keller, Emoji use in Twitter White Nationalism Communication, in: Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing, 2019, pp. 201–205.
[35] F. Rangel, P. Rosso, Overview of the 7th Author pPofiling 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.
[36] 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 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, 2020.
[37] R. F. S. Dias, I. Paraboni, Combined CNN+ RNN Bot and Gender Profiling., in: Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop, 2019.

[38] J. R. P. Fontcuberta, G. L. De la Peña Sarracén, Bots and Gender Profiling using a Deep Learning Approach., in: Working Notes Papers of the CLEF 2019 Evaluation Labs Volume 2380 of CEUR Workshop, 2019.
[39] Y. Joo, I. Hwang, L. Cappellato, N. Ferro, D. Losada, H. Müller, Author Profiling on Social Media: An Ensemble Learning Model Using Various Features, Notebook for PAN at CLEF (2019).
[40] J. Buda, F. Bolonyai, An Ensemble Model Using N-grams and Statistical Features to Identify Fake News Spreaders on Twitter, in: CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, 2020.
[41] J. Pizarro, Using N-grams to detect Fake News Spreaders on Twitter, in: CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, 2020.
[42] F. Rangel, P. Rosso, On the implications of the general data protection regulation on the organisation of evaluation tasks, Language and Law= Linguagem e Direito 5 (2019) 95–117.
[43] F. Rangel, P. Rosso, M. Franco-Salvador, A Low Dimensionality Representation for Language Variety Identification, in: In 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing'16. Springer-Verlag, LNCS(9624), 2018, pp. 156–169.
[44] T. Gollub, B. Stein, S. Burrows, Ousting ivory tower research: Towards a web framework for providing experiments as a service, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 1125–1126.
[45] T. Gollub, B. Stein, S. Burrows, D. Hoppe, Tira: Configuring, executing, and disseminating information retrieval experiments, in: 2012 23rd International Workshop on Database and Expert Systems Applications, IEEE, 2012, pp. 151–155.
[46] T. Gollub, M. Potthast, A. Beyer, M. Busse, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, Recent trends in digital text forensics and its evaluation, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2013, pp. 282–302.
[47] C. M. del Campo Rodriguez, G. Sidorov, I. Batyrshin, Hate Speech Detection on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[48] Álvaro Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, Profiling Hate Speech Spreaders on Twitter: Transformers and Mixed Pooling, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[49] I. B. Schlicht, A. F. M. de Paula, An Unified and Multi-lingual User Profiling for Detecting Haters, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[50] R. Jain, D. Goel, P. Sahu, Profiling Hate Speech Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[51] H. B. Giglou, T. Rahgooy, J. Razmara, M. Rahgouy, Z. Rahgooy, Profiling Haters on Twitter using Statistical and Contextualized Embeddings, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[52] I. Vogel, M. Meghana, Profiling Hate Speech Spreaders on Twitter: SVM vs. Bi-LSTM, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[53] J. Alcañiz, J. Andrés, Profiling Hate Spreaders using Word n-Grams, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[54] T. Anwar, Identify Hate Speech Spreaders on Twitter using Transformer Embeddings Features, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[55] C. Bagdon, Profiling Spreaders of Hate Speech with N-grams and RoBERTa, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[56] D. Espinosa, G. Sidorov, Profiling Hate Speech Spreaders using Characters and Words N-grams, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[57] H. Cabrera, S. Miranda, E. Tellez, INFOTEC-LaBD at PAN@CLEF21: Profiling Hate Speech Spreaders on Twitter through Emotion-based Representations, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[58] K. G. Das, B. G. Patra, Profiling Hate Speech Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[59] E. Puertas, J. C. Martinez-Santos, Phonetic Detection for Hate Speech Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[60] D. Dukić, A. S. Kržić, Detection of Hate Speech Spreaders with BERT, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[61] E. Finogeev, M. Kaprielova, A. Chashchin, K. Grashchenkov, G. Gorbachev, O. Bakhteev, Hate Speech Spreader Detection using Contextualized Word Embeddings, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[62] T. Ceron, C. Casula, Exploiting Contextualized Word Representations to Profile Haters on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[63] C. Ikae, UniNE at PAN-CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[64] R. J. Gallagher, M. R. Frank, L. Mitchell, A. J. Schwartz, A. J. Reagan, C. M. Danforth, P. S. Dodds, Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts, EPJ Data Science 10 (2021) 4.
[65] C. M. V. de Andrade, M. A. Gonçalves, Profiling Hate Speech Spreaders on Twitter: Textual Analysis of Tweets, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[66] M. Siino, E. D. Nuovo, I. Tinnirello, M. L. Cascia, Detection of Hate Speech Spreaders using

Convolutional Neural Networks, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[67] F. Balouchzahi, S. H. L., G. Sidorov, HSSD: Hate Speech Spreader Detection using N-grams and Voting Classifier, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[68] E. Katona, J. Buda, F. Bolonyai, Using n-Grams and Statistical Features and to Identify Hate Speech Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[69] R. Cervero, Use of Lexical and Psycho-Emotional Information to Detect Hate Speech Spreaders on Twitter - Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[70] I. Briggs Myers, P. B. Myers, Gifts differing: Understanding personality type, Davis-Black Publishing (1995).
[71] S. M. Mohammad, P. D. Turney, Nrc emotion lexicon, National Research Council, Canada 2 (2013).
[72] M. Lai, M. A. Stranisci, C. Bosco, R. Damiano, V. Patti, HaMor at the Profiling Hate Speech Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[73] D. Tosev, S. Gievska, Multi-level Stacked Ensemble with Sparse and Dense Features for Hate Speech Detection on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[74] L. G. Zaragozá, S. H. Pinto, Profiling Hate Speech Spreaders on Twitter using Stylistic Features and Word Embeddings, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[75] K. O. Akomeah, U. Kruschwitz, B. Ludwig, University of Regensburg @ PAN: Profiling Hate Speech Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[76] E. Alzahrani, L. Jololian, Profiling Hate Speech Spreaders by Classifying Micro Texts Using BERT Model, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[77] M. Uzan, Y. Hacohen-Kerner, Profiling Hate Speech Spreaders on Twitter using BERT and LSTM models, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[78] R. L. Tamayo, D. Castro-Castro, R. O. Bueno, Deep Modeling of Latent Representations for Twitter Profiles on Hate Speech Spreaders Identification. Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[79] D. Irani, A. Wrat, S. Amir, Social Media Hate Speech Detection with Learned User Representations, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[80] J. Höllig, Y. S. Lee, N. Seemann, M. Geierhos, Effective Detection of Hate Speech Spreaders on Twitter, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[81] P. P. F. de Mora, C. H. Barbera, Local Classification with Recurrent Neural Network, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.