

UniNE at PAN-CLEF 2021

(Notebook for PAN at CLEF 2021)

Catherine Ikae¹

¹University of Neuchâtel, Switzerland, Avenue du 1er-Mars 26, 2000 Neuchâtel, Switzerland

Abstract

The paper describes the work done on the PAN 2021 task about profiling Hate Speech Spreaders in Spanish and English messages extracted from Twitter. We implement a simple Ensemble Classifier class that allows us to combine seven different machine-learning classifiers, which predict a class by simply taking the majority rule of the predictions by the classifiers. We also propose a reduced set of features that are obtained by considering terms with $df > 3$ and $tf > 1$ thereby eliminating terms that only appear once in the corpus. The features are ranked according to their term difference in each category. Each category contributes an equal number of features to the classification task. With 800 features from each class, our model achieves an accuracy of 0.66 for the English dataset and 0.81 for the Spanish dataset attaining an average score of 0.735.

Keywords

Author profiling, Ensemble classifier, Two-step feature selection, profiling Hate Speech Spreaders

1. Introduction

Hate Speech is defined as any communication that disparages a person or a group based on some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics as defined in the Encyclopedia of the American Constitution [1]. Hate speech leads to discrimination against particular categories of people and undermines equality, which is a big issue for each civil society as explained by [2].

Given the large number of people using social media such as Twitter, Facebook as a means of communication and sharing ideas, which are of great benefit to humanity since information shared can reach a big audience in a short time. However, this benefit is not without challenges, these channels of communication have also been exploited to propagate hate speech and spread false news resulting in hate crimes [3]. The ubiquity of social networks and the low cost of using them render the propagation of hate speech a real concern for our society.

The lack of editorial control over the spread of hate speech has the potential to harm and damage the targeted members of the society. One can also add that the involved companies do not propose (or do not want to propose) a real control on the flow of information using their networks. Hence, a real need for an automatic mechanism to identify the presence of hate speech spreaders is an important research topic.


A large number of researchers have been drawn to this area to develop automated methods of hate speech detection [4]. This has become a major natural language processing (NLP) research

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

 catherine.ikae@unine.ch (C. Ikae)



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

 CEUR Workshop Proceedings (CEUR-WS.org)

topic in the recent years as, for example, [5] who suggest an evaluation task focusing on hate speech. Therefore, this study contributes to this area of research by applying the two-step feature selection technique [6] and an ensemble of ML classifiers on PAN-CLEF 2021 hate speech datasets.

2. Corpus: Overall Statistics

The training corpus was available in the English and Spanish languages. The English training dataset had 100 documents(authors) of label 0 (normal set of tweets) and 100 documents(authors) of label 1 (tweets containing some form of hate speech). The Spanish training dataset had 100 documents(authors) of label 0 and 100 documents(authors) of label 1. Each of these documents contained 200 tweets [7]. A document refers to a set of tweets by an author.

The mean length of tokens per English document with label 0 is 3,270, with the maximum number of tokens in a document being 4,231 and minimum is 2,057. Those with label 1 mean length of tokens per document is 3,303, with the maximum number of tokens in a document being 4,637 and minimum is 2,058.

For the Spanish documents, the mean length of tokens per document with label 0 is 3,198, With maximum number of tokens in a document being 4,256 and minimum is 2,134 and with label 1 is 3,565, with maximum number of tokens in a document being 4,246 and minimum is 2,374.

Table 1

Overall statistics about the training data in both languages

	English			Spanish	
	0	1		0	1
Nb. doc.	100	100	Nb. doc.	100	100
Nb tweets	20,000	20,000	Nb tweets	20,000	20,000
Mean length	3,270	3,303	Mean length	3,198	3,565
Voc	14,103	13,338	Voc	24,310	23,722

Example of tweets from each of the classes and languages are as shown in Table 2 and Table 3

3. Ensemble Classifier

An ensemble classifier is one that stacks several performing classifiers to come up with the best prediction from the combined classifiers [8]. The advantage of combining classifiers is to take advantage of the good performance from a set of classifiers that will result in better prediction results as compared to a single classifier on its own.

For our task, we consider the following classifiers namely:

1. Linear Discriminant Analysis (LDA) finds a linear combination of features that separates two or more classes of objects in order to classify them [9];
2. Gradient Boosting (GB) which modifies weak learners into strong learners [10];

Table 2

Sample of three tweets in English for each class

English	
Class 0	'First Catering 🤔 #URL#', 'RT #USER#: vibes with someone come naturally. don't force shit', 'RT #USER#: never force nobody to appreciate you.', 'RT #USER#: That camera pan to the horny niggas killing me', 'RT #USER#: Buying a \$50 dollar bottle of liquor for \$300 at a club is a wave I'll never hop on 🤔 pregame me pleaseee !', 'RT #USER#: If the casino had a uno table I'll tear some shit up 🤔🤔🤔🤔', 'RT #USER#: "Can you multitask?" yes actually i am losing my mind and chilling at the same time',
Class 1	RT #USER#: The four most dangerous people in the world today. #HASHTAG# #URL#', '#USER# #USER# And Clinton makes five.', 'RT #USER#: It really is a backward cult from the stone age, little to no respect for women. #URL#', 'RT #USER#: BREAKING: President Trump to host the Rush Limbaugh Show tomorrow as a Virtual MAGA Rally', "#USER# #USER# #USER# #USER# Who was that actor who said he was mugged by Trump supporters and wasn't? Smollet? Mollet?",

Table 3

Sample of three tweets in Spanish for each classes

Spanish	
Class 0	RT #USER#: hay pibes con uniforme en la calle???? parece una realidad paralela', '#USER# pero si vas a estar re linda, cara de verga', '#USER# viste, por eso es MI novia', 'estoy comiendo <u>sanguchitos</u> de salame y queso, no puedo ser mas feliz', 'RT #USER#: MESSIRVE es el mejor vocablo creado jamás, llegó un punto en el que se lo digo hasta a mi vieja, es impresionante',
Class 1	#USER# De todas formas tu consejo y comentario me hace reflexionar y lo tendré en cuenta. Muchas gracias', '#USER# A los que no vamos a invitar a votar es a tus amigos, colegas de los terroristas de ETA, a esos no...', '#USER# Yo me alegro de todo corazón ❤️❤️❤️', '#USER# #USER# #USER# #USER# Descrito a la perfección 🙌🙌🙌🙌',

3. Extremely Randomized Trees (Extra Trees ET) is an ensemble learning technique which aggregates the results of multiple de-correlated decision trees constructed from the original training sample to obtain its classification result [11];
4. Gaussian Naive Bayes (G_NB) a variant of Naive Bayes that follows Gaussian normal distribution [12];
5. Bernoulli Naive Bayes (B_NB) [12] performs classification by assuming each feature to be a binary-valued (Bernoulli, Boolean);
6. Random Forest (RF) an ensemble of decision trees that combines learning models to increases classification accuracy [13];
7. AdaBoost creates a strong classifier from a number of weak classifiers [14].

We used the scikit-learn Python machine learning library that provides an implementation of stacking for machine learning [8] to integrate these classifiers.

4. Feature Selection

Good classification results come from a good feature set generated from that training dataset. The features are also used to understand and explain the difference between the hate speech

spreaders and other users. We propose a technique that is capable of reducing the features space, the two-stage feature selection strategy [6].

The two-stage feature selection strategy works by considering tokens according to their document frequency (df) and term frequency difference. A threshold of three ($df > 3$) and ($tf > 1$) was used for this task. With these two constraints, we create a feature set capable of distinguishing each category. From the reduced number of tokens obtained by applying $df > 3$, a term frequency difference is computed but only tokens with a term frequency greater than 1 ($tf > 1$) are put into consideration to leave out those tokens that appear only once in the text.

With 70 documents taken from class 0 and 70 from class 1, we create our training set and the remaining 30 from class 0, 30 from class 1 is used as the test set. The features extracted from this selection is as described below: We begin our feature selection with 11723 features from class 0 and 11161 features from class 1. The features are reduced to 5970 for class 0 and 5799 for class 1 by considering only terms with $tf > 1$. The final reduced set is obtained by using frequency difference between the tokens from the two classes and checking if it has a $df > 3$. With the features ranked according to their term frequency difference in each class, the number of features used in creating the model can be selected in descending order from each category.

Table 4
Two-step feature selection for English and Spanish dataset

	English			Spanish		
	All	tf >1	tf diff and df >3	All	tf >1	tf diff and df >3
vocabulary_0	11723	5970	2932	19311	8135	3230
vocabulary_1	11161	5799	2855	19030	8519	3596
Total number of features			5787			6826

Using Shift Graphs [15], words are sorted by their absolute contribution to the difference between classes. The shift graphs are created with the document frequencies of the resulting features Figure 1 and Figure 2. With sys 2 belonging to class 1 and sys 1 to class 0. Words with high discriminating power in a class are shown at the top of the chart with longer bars and those with lower discriminating power have shorter bars. The bars represent the document frequency difference between classes. The same approach is used in the entire training dataset to obtain features that will be used to create the model for testing.

5. Evaluation

To train our model, features are extracted from the training documents by taking into account the steps explained in section 4. Features to be considered must have a $tf > 1$ and $df > 3$ from which the ranking is done according to the difference in term frequencies. The k feature set at each selection picks equal values from each subset, that is half the feature from class 0 and the other half from class 1.

The value of k is increased from 200 (100 from class 0 and 100 from class 1) to 2000. The accuracy of several classifiers are computed as shown in the table 5 and table 6. It was easy to analyse the performance of the classifiers where we can see that an increase in the number of features also had an increase in the accuracy of the classification. Taking an example of the

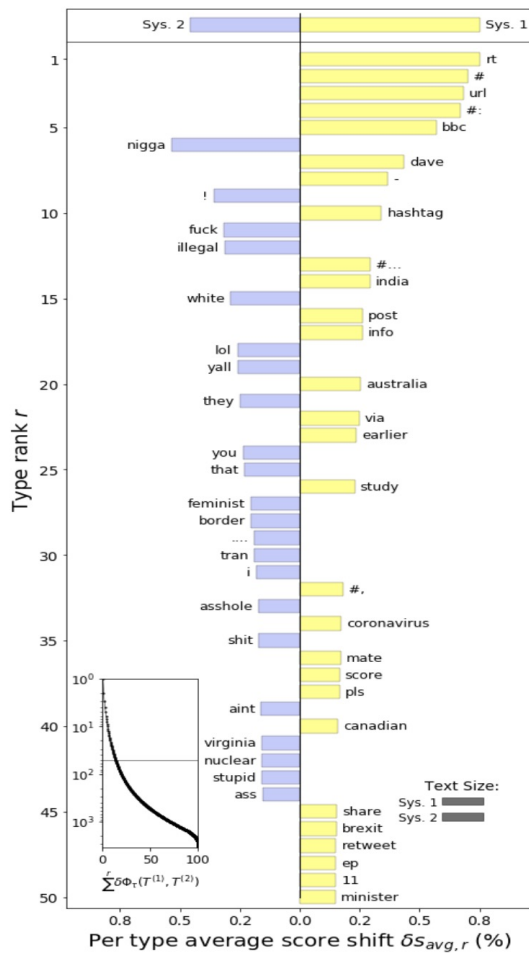


Figure 1: Shift graph for the English dataset

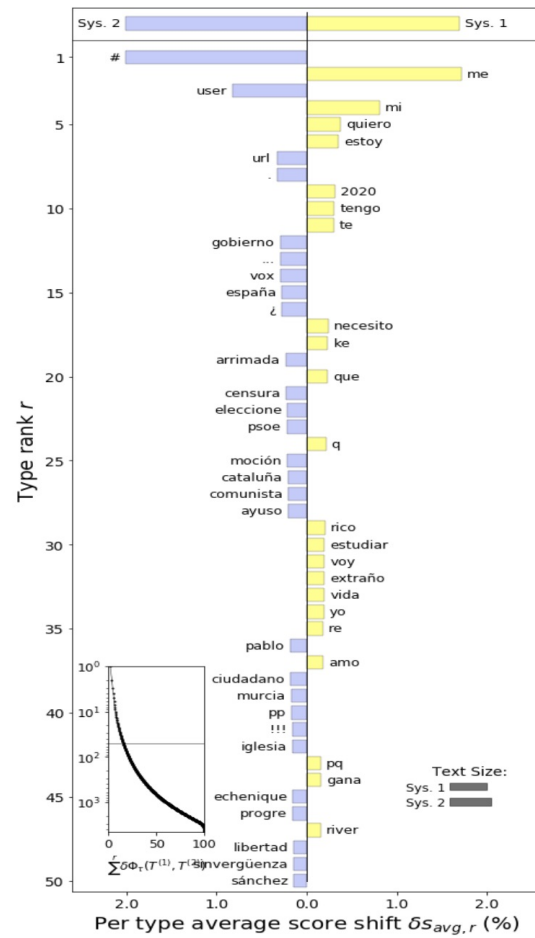


Figure 2: Shift graph for the Spanish dataset

LDA line one of the Table 5 and Table 6 shows increased accuracy from $k = 200$ to $k = 1600$ where we get a maximum accuracy of 0.82. An increase in the features at this point decreases the accuracy to 0.75 at $k = 2000$. Since not all classifiers produced similar results as the LDA, an ensemble of two classifiers was built with LDA and G_NB that gave an overall performance best accuracy at $k = 1600$ (800 from class 0 and 800 from class 1).

Table 7 depicts the accuracy rate achieved with our model under different conditions and for both languages. In the first row, 800 words from class 0 and 800 words from class 1 have been used to build the document surrogates and an ensemble of only two classifiers is used for classification giving an average score of 0.655. In the second line, the vocabulary size is kept the same but an ensemble of seven classifiers are used namely: Linear Discriminant Analysis, Gradient Boosting, Extremely Randomized Trees, Gaussian Naive Bayes, Bernoulli Naive Bayes, Random Forest and AdaBoost resulting into an average score on 0.735.

Table 5

Evaluation based on different feature sizes

	English									
Classifiers	200	400	600	800	1000	1200	1400	1600	1800	2000
LDA	0.62	0.55	0.57	0.58	0.60	0.60	0.60	0.65	0.68	0.65
GaussianPro	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
GradientBoost	0.65	0.55	0.60	0.65	0.58	0.53	0.48	0.48	0.47	0.48
ExtraTrees	0.68	0.60	0.58	0.58	0.57	0.67	0.55	0.68	0.58	0.63
KNN	0.52	0.48	0.50	0.48	0.50	0.48	0.50	0.50	0.52	0.52
GaussianNB	0.50	0.50	0.62	0.60	0.63	0.68	0.73	0.68	0.68	0.70
MultinomialNB	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
BernoulliNB	0.50	0.57	0.58	0.62	0.62	0.60	0.63	0.65	0.62	0.63
DecisionTree	0.60	0.50	0.57	0.47	0.50	0.57	0.57	0.58	0.55	0.55
RandomForest	0.65	0.53	0.63	0.55	0.58	0.60	0.50	0.55	0.65	0.60
LogisticReg	0.53	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
MLP	0.50	0.52	0.52	0.53	0.53	0.55	0.55	0.55	0.55	0.57
AdaBoost	0.58	0.53	0.50	0.55	0.60	0.52	0.52	0.57	0.57	0.57
Bagging	0.58	0.53	0.58	0.52	0.62	0.67	0.58	0.58	0.58	0.58
SGD	0.52	0.52	0.55	0.52	0.45	0.52	0.53	0.55	0.52	0.52
XGB	0.62	0.53	0.63	0.57	0.58	0.55	0.53	0.50	0.53	0.53
SVM	0.57	0.57	0.57	0.58	0.58	0.58	0.58	0.58	0.58	0.58
Ensemble(LDA + G_NB)	0.57	0.52	0.62	0.62	0.63	0.68	0.67	0.72	0.70	0.58

6. Conclusion

The paper describes the machine learning ensemble approach for hate speech spreaders detection task. We proposed an ensemble based on Linear Discriminant Analysis, Gradient Boosting, Extremely Randomized Trees, Gaussian Naive Bayes, Bernoulli Naive Bayes, Random Forest and AdaBoost classifiers. The resulting performance gave us an accuracy of about 0.66 for the English dataset and 0.81 for the Spanish dataset. Our approach is capable of distinguishing hate/non-hate speech spreaders since the features set used in the classification are drawn from both classes in equal numbers. A term frequency difference is used to determine the discriminating power of each feature in the class. These features indicate the difference that exists between the two classes and are ranked according to their frequency difference. For future work, the idea is to compare chi2 and mutual information feature ranking with the hope of boosting the feature selection.

Table 6

Evaluation based on different feature sizes

Spanish										
Classifiers	200	400	600	800	1000	1200	1400	1600	1800	2000
LDA	0.61	0.63	0.63	0.77	0.77	0.82	0.78	0.82	0.80	0.75
GaussianPro	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
GradientBoost	0.72	0.72	0.68	0.72	0.73	0.72	0.70	0.75	0.73	0.73
ExtraTrees	0.77	0.77	0.75	0.72	0.72	0.72	0.73	0.77	0.73	0.70
KNN	0.60	0.60	0.60	0.60	0.58	0.58	0.58	0.58	0.58	0.58
GaussianNB	0.72	0.73	0.73	0.75	0.78	0.78	0.72	0.77	0.75	0.75
MultinomialNB	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
BernoulliNB	0.72	0.72	0.72	0.70	0.72	0.72	0.73	0.72	0.70	0.70
DecisionTree	0.60	0.65	0.65	0.57	0.58	0.58	0.60	0.60	0.70	0.65
RandomForest	0.73	0.73	0.77	0.68	0.73	0.78	0.70	0.73	0.75	0.77
LogisticReg	0.58	0.58	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
MLP	0.63	0.63	0.65	0.65	0.65	0.65	0.65	0.65	0.67	0.67
AdaBoost	0.63	0.70	0.73	0.67	0.73	0.62	0.75	0.70	0.63	0.72
Bagging	0.65	0.63	0.65	0.72	0.70	0.68	0.72	0.73	0.65	0.70
SGD	0.63	0.47	0.58	0.48	0.48	0.50	0.57	0.57	0.55	0.60
XGB	0.68	0.72	0.73	0.75	0.75	0.77	0.75	0.77	0.73	0.72
SVM	0.58	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.58
Ensemble(LDA + G_NB)	0.68	0.68	0.70	0.80	0.78	0.82	0.76	0.82	0.78	0.73

Table 7

Official Evaluation with (k = 1600)

TIRA Test Results			
	ENGLISH	SPANISH	Average Score
Ensemble (LDA + G_NB)	0.57	0.74	0.655
Ensemble (LDA + G_NB + B_NB + GB + ET + RF + ADB)	0.66	0.81	0.735

References

- [1] Encyclopedia of the American Constitution, 2nd ed. / adam winkler, associate editor for the second edition. ed., Macmillan Reference USA, New York, 2000.
- [2] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1 – 30.
- [3] S. Abro, S. Shaikh, Z. H. Khand, Z. Ali, S. Khan, G. Mujtaba, Automatic hate speech detection using machine learning: A comparative study, International Journal of Advanced Computer Science and Applications 11 (2020). URL: <http://dx.doi.org/10.14569/IJACSA.2020.0110861>. doi:10.14569/IJACSA.2020.0110861.
- [4] 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, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: <https://www.aclweb.org/anthology/W17-1101>. doi:10.18653/v1/

W17-1101.

- [5] Ò. Garibo i Orts, Multilingual detection of hate speech against immigrants and women in Twitter at SemEval-2019 task 5: Frequency analysis interpolation for hate in speech detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 460–463. URL: <https://www.aclweb.org/anthology/S19-2081>. doi:10.18653/v1/S19-2081.
- [6] C. Ikae, S. Nath, J. Savoy, Unine at pan-clef 2019: Bots and gender task, in: CLEF, 2019.
- [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] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [9] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, *Linear Discriminant Analysis*, Springer New York, New York, NY, 2013, pp. 27–33. URL: https://doi.org/10.1007/978-1-4419-9878-1_4. doi:10.1007/978-1-4419-9878-1_4.
- [10] R. E. Schapire, The strength of weak learnability, in: *Machine Learning*, 1990.
- [11] P. Geurts, Extremely randomized trees, in: *MACHINE LEARNING*, 2003, p. 2006.
- [12] T. Ngo, *Data mining: Practical machine learning tools and technique*, third edition by ian h. witten, eibe frank, mark a. hell, *SIGSOFT Softw. Eng. Notes* 36 (2011) 51–52. URL: <https://doi.org/10.1145/2020976.2021004>. doi:10.1145/2020976.2021004.
- [13] L. Breiman, 1 random forests–random features, 1999.
- [14] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning – data mining, inference, and prediction*, ????
- [15] R. J. Gallagher, M. Frank, L. Mitchell, A. J. Schwartz, A. J. Reagan, C. Danforth, P. Dodds, Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts, *EPJ Data Science* 10 (2021) 1–29.