HSSD: Hate Speech Spreader Detection using N-grams and Voting Classifier

(Notebook for PAN at CLEF 2021)

Fazlourrahman Balouchzahi[1], Hosahalli Lakshmaiah Shashirekha[2] and Grigori Sidorov[1]

[1] Center for Computing Research, Instituto Politécnico Nacional, CDMX, Mexico

[2] Department of Computer Science, Mangalore University, Mangalore, India

Abstract

Profane or abusive speech with the intention of humiliating and targeting individuals, a specific community or groups of people is called Hate Speech (HS). Identifying and blocking HS contents is only a temporary solution. Instead, developing systems that are able to detect and profile the content polluters who share HS will be a better option. In this paper, we, team MUCIC, present the proposed Voting Classifier (VC) submitted to Hate Speech Spreader Detection shared task organized by PAN 2021. The task includes profiling HS spreaders for two languages, namely, English and Spanish from the text collected from Twitter. This task can be modeled as a binary text classification problem to classify an author (Twitter user) based on his/her tweets as 'Hate speech spreader' or 'Not'. The proposed models utilizes a combination of traditional char and word n-grams with syntactic ngrams as features extracted from the training set. These features are fed to a VC that employs three Machine Learning (ML) classifiers namely, Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) with hard and soft voting. The proposed models with accuracies of 73% and 83% for English and Spanish languages respectively, obtained second rank in the shared task.

Keywords

Hate Speech Spreader, Machine Learning, N-grams, Voting Classifier,

1. Introduction

Rapid dissemination, low cost, ease of access, and more importantly anonymity are the significant features of social media in current era [1, 2, 3]. There are so many religions, communities, groups of people and their subdivisions in this world whose thoughts and beliefs vary from one another. Mutual tolerance and respect is very essential for co-existence and peaceful living [4] on this earth. However, in some cases, one group's dogma can be against another as well creating panic and disturbances in the society. With inimical intentions or just for fun, there are users who share HS and profane content over social media or even offline. Online HS contents

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

frs1_b@yahoo.com (F. Balouchzahi); hlsrekha@gmail.com (H. L. Shashirekha); sidorov@cic.ipn.mx (G. Sidorov)

https://mangaloreuniversity.ac.in/dr-h-l-shashirekha (H. L. Shashirekha); http://www.cic.ipn.mx/~sidorov/ (G. Sidorov)

0000-0003-1937-3475 (F. Balouchzahi)

on social media are more fearsome and troublesome due to rapid dissemination of information [5]. HS contents usually originate from people or a group who are prejudiced with the intention of discriminating and targeting a race, religion or with sexual orientation of people who are noxious and harmful for society. Hence, the task of HS detection and profiling the spreaders is being indispensable [6, 7] in order to avoid the spread of HS and the possible damage it could cause to the society.

Appropriate tools and benchmarked labeled corpora are required to address the challenges of HS detection and profiling the spreaders [8]. In order to address these challenges, PAN [9] at Conference and Labs of the Evaluation Forum (CLEF) 2021 has called for a shared task: Profiling Hate Speech Spreaders on Twitter [10] for two languages namely, English and Spanish. The datasets provided by PAN consists of texts collected from Twitter and the task can be modeled as a binary Text Classification (TC) problem where a user based on his/her tweets can be identified as 'HS spreader' or 'Not'. As one of the participating team in this task, we, team MUCIC, have proposed an ensemble model that utilize the strength of three Machine Learning (ML) classifiers namely, Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) as estimators to build a robust VC.

LR is an impressive algorithm for binary and linear classification problems which models the probability of a discrete outcome from an input variable. Ease of realize and exquisite performance are the major features of this algorithm for binary classification [11]. SVM as a supervised ML algorithm has been widely used for classification and regression tasks. The main significance of SVM is identifying optimal boundary which effectively distinguish the classes in training data. SVM uses kernel trick technique to transform data and based on these transformations it will find an optimal boundary between the possible outputs . While a single Decision Tree (DT) consists of root and decision nodes with a top down greedy approach that splits the dataset into smaller subsets, RF is itself an ensemble learning model which employs a set of DTs and computes majority voting for the prediction of terminal nodes to determine the final prediction for the given input [12, 13].

Traditional n-grams are a set of co-occurring items or elements such as characters, words, Part-Of-Speech (POS) tags, etc. as they appear in a text. But, the idea of Syntactic n-grams (sn-grams) is to follow a path in the syntactic tree to construct n-grams, rather than taking them from surface representation. In other words, the sequence of words that appear in the path of a syntactic tree are considered as neighbors and the real neighbors of words based on syntactic relations [14, 15, 16] are extracted. To obtain the benefits of both the n-grams structures, the traditional char and word n-grams are extracted and combined with sn-grams as a feature set and transformed into vectors using CountVectorizer to feed the VC model. Rest of the paper is organized as follows: the related work and methodology are discussed in Section 2 and 3 respectively followed by results in Section 4. The paper eventually concludes with future work in Section 5.

2. Related Work

Most of HS detection tasks are modeled as short TC and rarely has been explored as profiling task. Some of the recent works on HS detection and text profiling have been reviewed here.

Zimmerman et al. [17] has proposed an ensemble of Deep Learning (DL) models for HS detection and also Sentiments Analysis (SA) from tweets. The authors ensembled 10 Convolution Neural Network (CNN) models by summing softmax results from the underlying models and then averaging it. Considering the average soft-max score of all models, the class with highest average is assigned to the given tweet. Utilizing the publicly available embedding models, this model was evaluated on two datasets, namely, abusive speech [18] and SemEval 2013 SA [19] and obtained average F1-scores of 77.83 and 70.36 respectively, with batch size and epochs of 10 each. HASOC 2020 [20] shared task organized by Forum for Information Retrieval Evaluation (FIRE) 2020 consists of two subtasks; i) a binary TC task where a given text should be categorized as HOF (containing HS contents) or NOT (Not Offensive) and ii) texts identified as HOF should be further classified into one of three categories namely, Hate speech (HATE), OFFENSIVE and PROFANITY. Datasets for this task has been provided for three languages, namely, English, Hindi, and German as detailed in [18]. Overall results reported by HASOC shows very competitive performances among the teams and differences between the F1-score of best performances and average ones are less than 0.04.

As a participant of HASOC 2020, Balouchzahi et al. [1] developed two models namely, ensemble of ML classifiers (LR, SVM, and RF) and Universal Language Model Fine-Tuning (ULMFiT) based on Transfer Learning approaches. The authors also employed ULMFiT as an estimator along with LR and RF. Texts are preprocessed by removing punctuations, stopwords, non-alphabets and unnecessary characters. fast.ai[1] and sklearn[2] libraries are used to build ULMFiT model and ML classifiers using pre-trained LM and combination of char and word n-grams respectively. For the first subtask an ensemble of SVM, LR, and ULMFiT obtained 0.497 and 0.518 F1-scores for English and Hindi respectively and ensemble of LR, SVM, and RF achieved 0.504 F1-score for German language. Also ULMFiT model submitted for second subtask in English language achieved F1-score of 0.265. Shashirekha et al. [21] ensembled three ML classifiers namely, Gradient Boosting, Random Forest and eXtreme Gradient Boosting as VC with soft voting configuration for HASOC 2020. After removing punctuation symbols, numeric data, stop words, uninformative words and frequently occurring words, features such as number of words, characters, punctuations, and length of the words are extracted from the training texts of all languages. Further, for English language, number of upper case characters, title words, and the frequency distribution of POS tags ie., Noun, Verb, Adjective, Adverb, and Pronoun are computed and used as additional features. These features are transformed to vectors using CountVectorizer and fed to the proposed model and obtained F1-scores of 0.5046, 0.5106, and 0.5033 for first subtask for English, German, and Hindi languages respectively. The proposed model also obtained 0.2596, 0.2595, and 0.2488 F1-scores for second subtask for English, German and Hindi respectively.

PAN at CLEF have managed to go further in identifying the content polluters who share HS, fake news, etc. or identifying bots from human followed by gender detection and profiling. Some of them are PAN 2018: Multimodal Gender Identification In Twitter [22], PAN 2019: Bots and Gender Profiling in Twitter [23] and PAN 2020: Profiling Fake News Spreaders on Twitter [24]. The task of profiling fake news spreaders on Twitter in PAN 2020 consists of datasets

[1] https://www.fast.ai/

[2] https://scikit-learn.org/stable/

for Spanish and English languages which includes 100 tweets per user and totally 300 users per language as training set and 100 tweets per user and totally 100 users per language as test set. Shashirekha et al. [2, 3] submitted two models, namely, ULMFiT and ensemble of ML classifiers as a VC for this task. They scraped raw texts from Wikipedia for Spanish and English languages and applied basic preprocessing steps. Preprocessed texts were used to train general domain Language Model (LM) and texts from training set were used to fine-tune the LM and finally the LM was employed to build target model for detecting fake news spreaders. Similar to Balouchzahi et al. [1] fast.ai library has been used to build LM and target model. For ML VC model construction, training set was first preprocessed by eliminating stopwords and punctuation, converting emoji to text and lemmatizing the words followed by feature extraction. Unigram TFIDF, N-gram TF combined with Doc2vec are extracted as features and scaled by MaxAbsScaler. A combination of Chi-square test, Mutual Information, and F-test algorithms are used to select important features which are in turn used to train the proposed VC. As per the results reported by PAN, ULMFiT and ML VC models obtained average accuracies of 0.63 and 0.70 respectively.

3. Methodology

The significance of ensembling ML models lies in improving the strength and covering the weakness of individual classifier models. Taking a note of this concept, a VC model of three ML estimators namely, SVM, LR, and RF is developed by exploiting hard and soft voting configuration for English and Spanish languages respectively.

ML models used in the proposed VC are chosen because of their efficient performances for binary classification as proved in the available literature and based on our experiments. While RF which is already a method of ensembling utilize 10,000 decision trees as estimators, SVM uses linear kernel. Rest of parameters for these two models and all parameters of LR estimator have been set to default. As a preprocessing step, texts are striped and hashtags such as USER, URL, and RT are removed and all words are converted to lower case for English. However, preprocessing is avoided for Spanish language texts as our experiments without preprocessing performed better.

A feature extraction module as shown in Figure 1 is used to extract char (2, 3, 4, 5) and word (2, 3) n-grams and sn-grams (2, 3). SNgramExtractor[3] library has been used to extract sn-grams from English and Spanish texts. The extracted features are transformed to vectors using CountVectorizer. Figure 2 illustrates the structure of proposed VC model graphically.

4. Experimental Results

4.1. Dataset

Datasets provided by PAN consists of a training set of 200 XML files for each language and each XML file represents a user with 200 tweets. The test set consists of 100 XML files per language and the proposed models should identify whether a user (represented by an XML file) is a ‘HS

[3] https://pypi.org/project/SNgramExtractor/

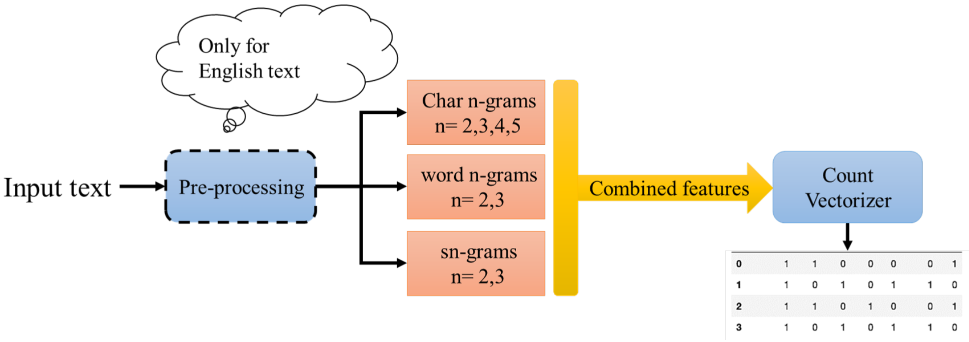


Figure 1: Feature Extraction module

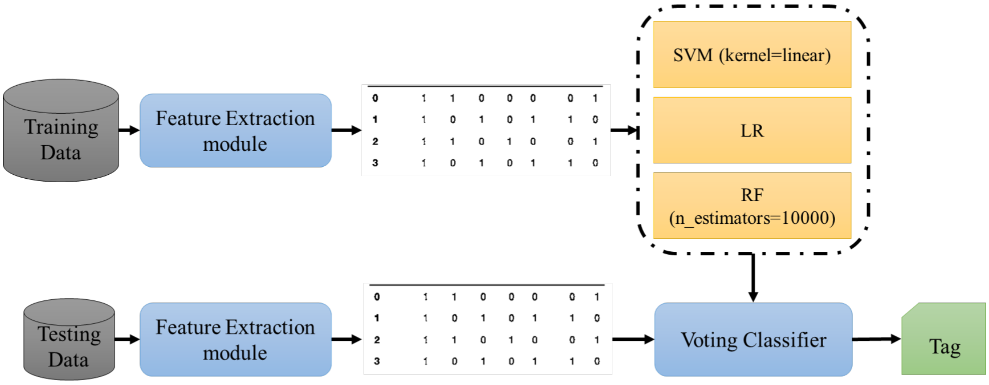


Figure 2: Structure of proposed VC model

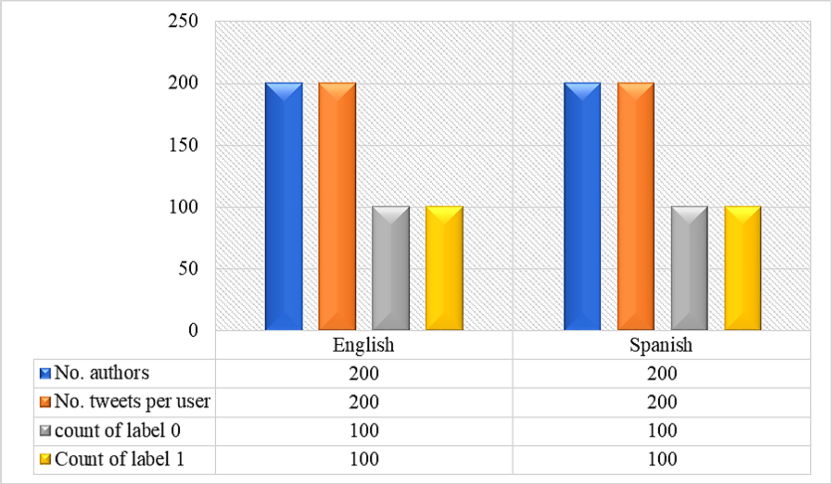
spreader' or 'Not' based on the analysis of tweets given in the XML files. Details of the training data along with label distribution which is presented in Figure 3 illustrates that the dataset is completely balanced.

4.2. Results

PAN uses TIRA Integrated Research Architecture submission system [25] that provides Virtual Machine (VM) for the shared task participants through which they can submit and evaluate their proposed models. As PAN encourages early bird submission of the models, the initial models of the proposed approach are submitted through TIRA and due to technical issues the final model and predictions on test set (labels) are submitted through mail. Performances of the models are evaluated by the task organizer based on accuracy metric and the results in shared task website[4] illustrate that the VC model obtained accuracies of 83% and 73% for Spanish and English texts respectively.

Performances of the best teams presented in Table 1 shows very competitive results and our proposed models (mentioned as MUCIC) obtained second rank in the shared task. The highest

[4] https://pan.webis.de/clef21/pan21-web/author-profiling.htmlresults



	English	Spanish
🟦 No. authors	200	200
🟧 No. tweets per user	200	200
⬜ count of label 0	100	100
🟨 Count of label 1	100	100

Figure 3: Distribution of labels in the training data provided by PAN

Table 1
Best performing teams in shared task

Team	English	Spanish	Average
SiinoDiNuovo	73.0	85.0	79.0
MUCIC	73.0	83.0	78.0
tamayo	74.0	82.0	78.0
andujar	72.0	82.0	77.0
anitei	72.0	82.0	77.0
anwar	72.0	82.0	77.0

accuracies reported for Spanish and English texts are 85% and 74% respectively.

5. Conclusion and Future Work

Following the adventures in text processing tasks, PAN 2021 called for a shared task to detect Hate Speech Spreaders in English and Spanish language tweets. This challenge is tackled by team MUCIC by building a robust VC using ML classifiers and traditional char and word n-grams along with syntactic n-grams as features to train VC model. Our team (MUCIC) obtained second rank with an average accuracy of 78% in the shared task. As future work we would like to explore more feature sets with ML models and also experimenting DL and TL approaches.

6. Acknowledgment

Team MUCIC deeply appreciates the efforts, guidance and support of the shared task organizers and reviewers for the valuable comments and suggestions.

References

[1] F. Balouchzahi, H. L. Shashirekha, Las for HASOC - learning approaches for hate speech and offensive content identification, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 145–151. URL: http://ceur-ws.org/Vol-2826/T2-6.pdf.

[2] H. L. Shashirekha, F. Balouchzahi, Ulmfit for twitter fake news spreader profiling, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_126.pdf.

[3] H. L. Shashirekha, M. D. Anusha, N. S. Prakash, Ensemble model for profiling fake news spreaders on twitter, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_136.pdf.

[4] V. Sinha, Theorising'talk'about'religious pluralism'and'religious harmony'in singapore, Journal of Contemporary Religion 20 (2005) 25–40.

[5] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, Overview of the evalita 2018 hate speech detection task, in: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, volume 2263, CEUR, 2018, pp. 1–9.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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: A. J. M. M. F. P. Guglielmo Faggioli, Nicola Ferro (Ed.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[11] A. Subasi, Practical Machine Learning for Data Analysis Using Python, Academic Press, 2020.

[12] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer Science & Business Media, 2009.

[13] K. Kirasich, T. Smith, B. Sadler, Random forest vs logistic regression: binary classification

for heterogeneous datasets, SMU Data Science Review 1 (2018) 9.

[14] G. Sidorov, Continuous and noncontinuous syntactic n-grams, in: Syntactic n-grams in Computational Linguistics, Springer, 2019, pp. 63–67.

[15] G. Sidorov, Syntactic dependency based n-grams in rule based automatic english as second language grammar correction, International Journal of Computational Linguistics and Applications 4 (2013) 169–188.

[16] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, L. Chanona-Hernández, Syntactic n-grams as machine learning features for natural language processing, Expert Systems with Applications 41 (2014) 853–860.

[17] 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.

[18] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[19] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, T. Wilson, SemEval-2013 task 2: Sentiment analysis in Twitter, in: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, Georgia, USA, 2013, pp. 312–320. URL: https://www.aclweb.org/anthology/S13-2052.

[20] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, 2020, pp. 29–32.

[21] M. D. Anusha, H. L. Shashirekha, An ensemble model for hate speech and offensive content identification in indo-european languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 253–259. URL: http://ceur-ws.org/Vol-2826/T2-20.pdf.

[22] F. Rangel, P. Rosso, M. Montes-y Gómez, M. Potthast, B. Stein, Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter, Working Notes Papers of the CLEF (2018) 1–38.

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

[24] 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.

[25] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, Tira integrated research architecture, in: Information Retrieval Evaluation in a Changing World, Springer, 2019, pp. 123–160.