IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Online Abuse Detection Using Machine Learning

[1]M.Sri Vidya [2]R.Haneesh, [3]K.Vinay, [4]Ashsih Biradar
[1]Assistant Professor, [2]Student, [3]Student, [4]Student
[1]Information Technology
[1]Osmani University, Hyderabad, India

Abstract : Hate Speech (HS) in social media such as Twitter is a complex phenomenon that attracted a significant body of research in the NLP. HS Spreaders (haters) aim to spread HS via social media. In this task, we aim to identify such haters. On one hand, our proposed class-dependent LDSE representation is fed to a linear SVM classifier to identify the haters based on general commonalities. On the other hand, stylistic features of individuals are captured by using extractive summarization of the tweets in conjunction with RoBERTa embedding before classifying them using another linear SVM classifier. Experimental results expressed as accuracies 0.67 and 0.80 over English and Spanish test sets respectively show efficacy of our approach in identifying the haters across different languages.

1. Introduction

The social medial platform enables millions to publicly share user-generted content. Regardless of different content types, a critical point of these platforms, such as Twitter, Facebook, YouTube, and Instagram, is that users can discuss content. Unfortunately, any user engaging online will always facing the risk of being targeted or harassed via abusive language, hatred expressed in the form of racism or sexism, with possible impact on his/her and the community in general. The challenge of creating effective policies to identify and appropriately respond to harassment is compounded by the difficulty of studying the phenomena at scale. Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics.To this end, in the Author Profiling task [1], we aim at identifying possible hate speech spread-0 ers (haters) on Twitter as a first step towards preventing hate speech from being propagated0 among online users. Also,this task runs based on a multilingual perspective for English and Spanish languages.

The rest of the paper is organized as follows. Section 2 presents related works. Section 30 describes the proposed method. Section 4 describes the dataset, experiments and discusses the obtained results. Finally, section 5 presents our conclusions.

2. RelatedWorks

The authors in [2] concluded that word embedding models like GloVe[3] and Word2Vec[4] are although widely used for toxic comment classification are in fact unable to handle out-of-vocabulary problem properly. However, word embeddings like FastText[5] were particularly suited for this task since it uses subword embedding. The ability to cope with unknown words is the reason why previous findings on the inferiority of word embeddings in comparison to word n-grams have become outdated. After introducing contextualized word embeddings such as BERT[6], some of the limitations of all previous word embeddings are resolved. In SemEval-2021 at the Toxic Span Detection task [7] used GloVe, GPT-2[8] and RoBERTa[9] to create empowered representation to overcome the out-of-vocabulary issue in the representation. They use the analogy that hateful words may never occur in the training time of the word embeddings like GloVe or Word2Vec, but however they are useful when they are combined with other representations since concatenations of them with LM models could create awareness signal to model.

3. ProposedMethod

In this section, we describe the details of our proposed model. Our proposed approach aims to predict whether the user is keen to spread hate speech or not.

proposed method. We used two different representations namely, contextualized representation using RoBERTa embedding, and character-based lower dimensionality statistical embedding (Char-LDSE) [16]. In the final, representations are fed into the classification modules. In the following, we described each component in detail.

3.1. ContextualizedRepresentation

The user content in this task consists of 200 tweets. To capture the user preferences in tweets and reduce the number of tweets to feed the word embeddings, we applied extractive text summarization. Next, we preprocessed the summary to feed RoBERTa embedding for creating a contextualized representation of user preferences. In the end, for the users, we had a representation of $\times$ 768 to feed the classification module, where 768 is the vectorized rep-resentation of summaries using word embedding. In the following, we describe summarization,preprocessing, and word embedding components.

3.1.1. Extractive Text Summarization:

It involves selecting phrases and sentences from the original text and including it in the final summary. We used Gensim [17], a Python library to summarize user tweets with summary ratio of **0.1** (selecting 20 tweets from user 200 tweets). The Gensim uses ***TextRank*** algorithm, which is based on ***PageRank*** algorithm for ranking search results.

1. Pre-process the given tweets (a built-in preprocessing in gensim).
2. Make a graph with tweets that are the vertices.
3. The graph has edges denoting the similarity between the two tweets at the vertices.
4. Run PageRank algorithm on this weighted graph.
5. Pick the highest-scoring vertices and append them to the summary.
6. Based on the ratio or the word count, the $_0$number of vertices to be picked is decided.

3.1.2. Preprocessing:

The preprocessing consists removal of URLs, hashtags, mentions, reserved words (RT, FAV), emojis, smileys, punchuations, special characters, and numbers from tweets. Speficicaly for URLs, hashtags, and mentions the following masked tags, #USER#, #URL#, #HASHTAG#.

3.1.3. RoBERTa:

It is an optimized version of BERT model. It builds on BERT's language masking strategy. RoBERTa modifies key hyperparameters in BERT, including removing BERT's next sentence pretraining objective and training with much larger mini-batches and learning rates. For the Egnlish language, we used the English version of RoBERTa base model, and for Spanish, we used SpanBERTa1 which is of the same size as BERT-Base and is trained on 18 GB of OSCAR's Spanish corpus.

3.2. Char-LDSE Representation

Char-LDSE[16] representation is applied to capture the stylistic features of user tweets and the probability of term occurrences in hate and none-hate spreaders. First, preprocessing is applied to user tweets; next, a character n-gram matrix with TFIDF weight is created. TFIDF matrix is utilized to calculate the LDSE. Next, the weighted probability of terms per class was obtained.

As a result, 0 and 1 embeddings are calculated for class 0 and class 1, respectively. In the end, using these embeddings, we calculated a matrix of×104 per class to feed classification models. Where 104 is the vectorzied distribution of user weighted probabilities of terms in each class (for class 0, 0 and for class 1 1 was utilized separately)

3.2.1. Preprocessing:

The preprocessing consists removal of special characters and localization of the tweets.

[1]https://github.com/chriskhanhtran/spanish-bert

3.2.2. TFIDF:

We apply the TFIDF weighting on the terms of the user tweets in the training set. We utilized charachter n-grams. Specifically, for English, we used range (2, 3), and for Spanish, we used a range of (3, 4). These ranges are obtained using a manual search. As a result, we obtain the following matrix.

10 10 ... 10 (0
0
))

1 2 ... 2 (0
0

...2 ...2 ... ...0)
= (12)1 (22)2 ... (2) (01)
– –
–
(1)1 (1)2 ... 1) (01)
– 1 – 2 .... ((
–

Where each row in the matrix TFIDF represents a user , each column represents vocabulary
term andrepresents its TFIDF weight and represents the assigned class (0 - class 0,
1
- class 1) of the user tweets. Also, and represent the number of the training set (users) and vocabulary size, respectively.

3.2.3. Lower Dimensionality Statistical Embedding (LDSE):

First, using matrix TFIDF, we obtain the class-dependent term weight embedding LDSE. This embedding contains the weights of each term for each class based on the following formulation.

∑)/ = (

∈ ((,) = ∑ ( )
)
∈

Next, we calculated LDSE for each class:

0

,, {0,1} 0
∀ ∈ ∈

0= 0(,), 0
∀ ∈

1= (,), 1
∀ ∈

3.2.4. Final Representation:

At the end, We employed the class-dependent LDSE, 0 and 1 to extract the final represen- tation of user tweets as follows for each class seperately:

1= (0), 2= (1)

Where ( ) contains the set of features showed in the followings and described in Table 1_0

(
)=,,,,,,,..., 100
{ 123}

Table 1
Set of features for each class (hater and none-hater)

avg	Theaverageweightof (,)fromausercontent
min	Theminimumweightof (,)fromausercontent
std	Thestandarddeviationoftheweightof (,)
prob	Theoverallweightof (,)fromausercontentdevidedbytotalnumberofterms | 1,..., 100 calculatingtheQ-thquantileofthedata.

Table 2
Number of authors in the PAN-AP-21 corpus created for profiling hate spreaders on Twitter.

Language	Traning	Testing	Total
English	200	100	300
Spanish	200	100	300

3.3. Classification Modules

There are many different types of ensembles; voting is one of them. It is one of the more general types. Voting involves training a learning algorithm to combine the predictions of several other learning algorithms. We used voting with the hard scheme, in which we trained three different classifiers with three different representations each. The Linear SVM with = 0.1 was utilized as a classifier algorithm for each representation that we obtained as described in Table 3 column Representation + Model.

4. Experiments and Results

In this section, we described the Author Profiling dataset. Next, we presented experimental results on the training set. Finally, we presented the proposed model's final results on the test set.

0

4.1. Dataset

Table 2 presents the statistics of the corpus that consists of 300 authors for each of the two languages, English and Spanish. For each author at least 200 Tweets collected. The corpus for each language is balanced, with 150 authors for each class (hater and none-hater spreaders). Dataset have splited into training and test sets, following the 66/34 proportion.

4.2. Experimental Results

We conducted a few experiments with Linear SVM and different representations. We mainly focused on 5-fold cross-validation mean accuracy. According to experiments for English, ensemble modeling with different representations outperforms single representation modeling.
Even in some cases (in Fold-4 and Fold-5), it gains higher accuracy than individuals. It means even contextualized representation which performs week in overall with a mean accuracy of 0.625, contributes to the ensemble model in a positive way.

Table 3

5-Fold Cross Validation Results. In the table Rep1 is Char-LDSE (0) and Rep2 is Char-LDSE (1) Lan Representation+Model Fold-1 Fold-2 Fold-3 Fold-4 Fold-5 MeanAccuracy

RoBERTa+LinearSVM 0.600 0.575 **0.675** 0.600 0.675 0.625 Rep1+LinearSVM **0.700 0.625** 0.650 0.650 0.750 0.675 en

Rep2+LinearSVM 0.650 0.575 0.600 0.650 0.800 0.655

Ensemble **0.700 0.625** 0.650 **0.675 0.825 0.695**

RoBERTa+LinearSVM 0.650 0.725 0.700 0.525 0.650 0.650 Rep1+LinearSVM 0.700 **0.825** 0.750 0.750 0.775 **0.760** es

Rep2+LinearSVM 0.675 0.800 0.750 0.750 0.775 0.750

Ensemble 0.675 **0.825** 0.750 0.750 0.775 **0.755**

Table 4

Submission Results

Language	Early Bird Submittion**0**	Final Submission0	Best
English	**0.670**	0.650	0.67
Spanish	0.730	**0.800**	0
Average	0.700	0.725	0.80
			0
			0.73

Spanish regarding the fact that the second model (Rep1:LDSE(0) + Linear**5** SVM) performs in a similar way that Ensemble works. Consequently, we relied upon the power of the ensemble approach for the final submission.

For early bird submission, we simply used XGBoost classifier with Char n-gram with TFIDF weights for the Spanish language. Howerver, we employed an SVM classifier with RBF kernel and LDSE representation with the tree-based feature selection model for English.

We particularly didn't rely on single representations because of the ambiguity of the user behaviors since hate spreaders may have multiple none-hate tweets in their own tweets too.

4.3. Final Evaluation

Following the previous results, for the final evaluation at TIRA platform [18], we applied statistical and contextual representations via ensemble models for hate speech spreaders detection. The obtained accuracy results for the final evaluation were as follows: in Spanish, **0.800**; in English, **0.670**; and **0.735** for both tasks. The official results are shown in Table 4 for early birds and final evaluation. We gained a better result for English at the early bird evaluation. However, for Spanish, we achieved higher accuracy at the final evaluation. In the final evaluation metrics,

The best scores of the submissions between the early birds and final submissions of each participant and each language have been considered. This means that in our case, we achieved the best score for English in early bird and the best score for Spanish in the final submission, so, overall achieved accuracy is 0.735.

5. Conclusion

In this paper, we proposed a model for Profiling Hate Speech Spreaders on the Twitter task in PAN 2021. We presented statistical and contextual representations via an ensemble approach for hate speech spreaders detection. In the final, we achieved an average accuracy of 0.735. Based on our manual evaluation, our approach is very capable of distinguishing hate/none-hate speech spreaders. The proposed algorithm implemented in Python and published on GitHub2 repository for research community.

6. References

- F. Rangel, G. L. D. L. P. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech0 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.
- J. Risch, R. Krestel, Toxic comment detection in online discussions, in: Deep Learning-Based Approaches for Sentiment Analysis, Springer, 2020, pp. 85–109.
- J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in:0
- ○ Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. URL:
- http://www.aclweb.org/anthology/D14-1162.
- T.Mikolov,K.Chen,G.Corrado,J.Dean,Efficientestimationofwordrepresentationsin vector space, 2013. arXiv:1301.3781.

- P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).
- J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
- 2021 . arXiv:2104 . 13164 .
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).
- Y.Liu,M.Ott,N.Goyal,J.Du,M.Joshi,D.Chen,O.Levy,M.Lewis,L.Zettlemoyer,V.Stoy-0 anov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.
- B. van Aken, J. Risch, R. Krestel, A. Löser, Challenges for toxic comment classification: An0
  - in-depth error analysis, in: Proceedings of the 2nd Workshop on Abusive Language Online0 (ALW2), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 33–42. URL: https://www.aclweb.org/anthology/W18-5105. doi:10.18653/v1/W18-5105.
- P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection0
  - in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.
- F.Rangel,A.Giachanou,B.Ghanem,P.Rosso, Overviewofthe8thAuthorProfilingTaskat0 PAN 2020: Profiling Fake News Spreaders on Twitter, in: L. Cappellato, C. Eickhoff, N.Ferro.