Detecting Hate Speech on English and Indo-Aryan Languages with BERT and Ensemble learning

Camilo Caparrós-Laiz[1], José Antonio García-Díaz[1] and Rafael Valencia-García[1]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain*

Abstract

The increasing use of social media platforms is making possible the communication between people around the world, including those with conflicting ideologies and cultures. However, some people use offensive language instead of having a polite conversation either because of little education or with the intention of intoxicating the debate. In this paper we analyze the results achieved by the UMUTeam of applying BERT models, either separate or combined with other popular models, for the HASOC'2021 shared-task for identifying offensive language in English, Hindi and Marathi. Our bests results are achieved with BERT for English classification subtask (1A), in which we reached a macro F1-score of 80.13%, and with ensemble learning for the rest of the subtasks, reaching macro F1-scores of 62.89% in English (subtask 1B), 75.20% and 51.67% in Hindi (subtasks 1A and 1B, respectively), and 84.23% in Marathi.

Keywords

Hate-speech detection, Feature engineering, Transformers, Low-resource languages, Deep-learning

1. Introduction

Social media platforms are places where people can freely communicate. However, the anonymity and remote communication makes it easier to have impolite, offensive or even hateful conversations. Although these platforms may have their own rules, heated arguments can go unnoticed or be wrongly censored. One way to avoid censorship is to provide tools for tagging offensive messages is by using methods that automatically detect offensive language. The biggest challenge regarding offensive-speech is the ability to interpret language. Naive methods may be based on the identification certain offensive keywords but do not take into account the context around the keyword. Modern approaches based on transformers are capable of detecting the context of the words, improving the accuracy of these kind of systems.

HASOC 2021's shared-task [1] offered two subtasks. The first subtask is divided into two: a binary classification problem to identify whether a post contains hate, offensive or profane content [2, 3] (1A) and a multi-classification problem to discriminate between hate, profane and offensive posts (1B). The second subtask is the identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL). The dataset for these tasks contains tweets which may or may

Forum for Information Retrieval Evaluation, December 13-17, 2021, India

camilo.caparrosl@um.es (C. Caparrós-Laiz); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

0000-0002-5191-7500 (C. Caparrós-Laiz); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

not be offensive in English, Hindi and Marathi. The UMUTeam only participated in subtasks 1A and 1B in all the proposed languages. Our approach to solve these tasks is by using state of the art models such as [4] BERT, which is able to capture language features, and combine it with other linguistic features (LF) extracted from UMUTextStats [5, 6] and by applying ensemble learning approaches.

We can find in the bibliography several works regarding hate-speech and offensive-speech identification. For example, in [7] the authors evaluated multiple models such as logistic regression, naive Bayes, decision trees and linear support vector machines (SVM). They found that logistic regression with L2 regularization by using 5-fold cross-validation and parameter grid search achieved the best results, with a precision of 91% and recall of 90% for offensive language, and a precision of 44% and recall of 61% for hate speech. Another related research is [8], in which the authors used a Multi-view SVM to classify hate speech. They fit different text features to a linear SVM and, finally, combine those features with a linear SVM. This model achieved a 80% of accuracy on Stormfront dataset and a 61% of accuracy on TRAC dataset. One benefit of the multi-view SVM approach is that it allows for some interpretability by identifying which classifier contributes the most. Other works have focused on specific types of hate-speech identification, such as misogyny. For example, in [6] the authors compiled and evaluated a dataset in Spanish regarding misogyny. They focused on determining violence against relevant women and also they analyzed differences between posts from European-Spanish and Latin American Spanish.

Most of the datasets are in English. However, many other languages do not have the same availability. Relevant work has been done to improve the accuracy on certain languages by using cross-lingual transfer learning as in [9] for Bengali, Hindi and Spanish, and as in [10] for German.

2. Methodology

We participate with three runs for the subtasks: 1A, 1B (English), 1A, 1B (Hindi), and 1A (Marathi). Due to lack of time, we could not complete the first run for subtasks 1B (Hindi) and 1A (Marathi).

For the first run, we use BERT [4] with HuggingFace [1] transformers library [11]. This library provides an automatic framework to train and predict a wide variety of models for supervised learning. The tokenizer and BERT model for English was `bert-base-uncased` and `bert-base-multilingual-uncased` for Hindi and Marathi. The BERT model is `BertForSequenceClassification`, which is used to classify sequences like text. We divided each dataset into training and validation to let us know beforehand the performance of the models.

For the second run, we combine the same neural network BERT with stylometric linguistic features [5, 6]. For this, we extract from BERT the encoding of the `[CLS]` token, as suggested in [12]. Next, we use Keras to combine the features and evaluate a total of 110 neural network models in which different number of hidden layers, neurons, learning rates, batch sizes, and activation functions. We ranked all the evaluated models with the macro F1-score over the validation split and select the one that achieved the best result.

[1] HuggingFace: https://huggingface.co/

For the third run, we prepare an ensemble of deep-learning classifiers based on the weighted mode. To calculate the importance of each classifier in the final output, we ranked each model based on the F1-score with the validation dataset. The involved neural networks models are the following: (1) linguistic features, (2) sentence embeddings from fastText [2] in English [13] and in Hindi and Marathi [14], (3) word embeddings from gloVe and fastText evaluated with convolutional and recurrent neural networks, and (4) BERT (as described for the first and second run). For each feature set, we evaluated 110 neural network models in order to decide what are the best hyperparameters.

3. Results

In this section, we show and discuss the results obtained by our methods. In each table, we present the three best methods of the overall results and our three best methods (UMUTeam) sorted by their macro F1-score obtained by the official HASOC leaderboard. Each table is a subtask (see Tables 1 and 2 for English, Tables Table 3 and 4 for Hindi, and Table 5 for Marathi).

Table 1
Results over English subtask 1A

Team	Method	Macro F1
NLP-CIC		**0.8305**
HUNLP		0.8215
neuro-utmn-thales		0.8199
UMUTeam	Run 1 - BERT	**0.8013**
UMUTeam	Run 3 - Ensemble	0.7959
UMUTeam	Run 2 - BERT+LF	0.7933

Table 2
Results over English subtask 1B

Team	Method	Macro F1
NLP-CIC		**0.6657**
neuro-utm-thales		0.6577
HASOC21rub		0.6482
UMUTeam	Run 3 - Ensemble	**0.6289**
UMUTeam	Run 2 - BERT+LF	0.6219
UMUTeam	Run 1 - BERT	0.3751

Our best result for the English subtask 1A is 80.13%, achieved by the BERT model. This result is very close to the best overall result (83.05%). The ensemble and LF models also had results close to the top results with 79.59% and 79.33%, respectively. For the English subtask 1B we achieved 62.89% with the ensemble model. Again, this result is very close to the top result

[2]fastText: https://fasttext.cc/

Table 3
Results over Hindi subtask 1A

Team	Method	Macro F1
t1		**0.7825**
Super Mario		0.7797
Hasnuhana		0.7797
UMUTeam	Run 3 - Ensemble	**0.7520**
UMUTeam	Run 1 - BERT	0.7185
UMUTeam	Run 2 - BERT+LF	0.6724

Table 4
Results over Hindi subtask 1B

Team	Method	Macro F1
NeuralSpace		**0.5603**
SATLab		0.5586
hate-busters		0.5582
UMUTeam	Run 2 - Ensemble	**0.5167**
UMUTeam	Run 1 - BERT+LF	0.4889

Table 5
Results over Marathi subtask 1A

Team	Method	Macro F1
WLV-RIT		**0.9144**
neuro-utmn-thales		0.8808
Hasnuhana		0.8756
UMUTeam	Run 2 - Ensemble	**0.8423**
UMUTeam	Run 1 - BERT+LF	0.8402

(66.57%). Our second run achieved similar results (62.19%). However, BERT got very bad results (37.51%). The best model for the Hindi subtask 1A achieved a macro F1-score of 75.20% with the ensemble model, 71.85% for BERT model and 67.24% for LF model. The results for Hindi subtask 1B are 51.67% for ensemble model and 48.89% for LF model. Lastly, for the Marathi subtask 1A we achieved a macro F1-score of 84.23% with the ensemble model and 84.02% with the LF model.

We notice that, generally, the ensemble learning model is the one that performs the best in all tasks except for the English subtask 1A in which BERT performs better. We can also observe that the English subtask 1A is quite near to top performing methods of the overall results. This finding suggests that the feature sets and the neural network models are complementary as their combination outperforms the results achieved by BERT.

We achieved near to state of the art results in English subtasks and decent results for Hindi and Marathi. We see that transformer models such as BERT are suitable for English tasks about hate speech and offensive language detection. However, other languages such as Hindi, which

is quite different from English, seem to perform worse. This may be for two reasons: (1) the model itself may not fit as well as other languages [15, 16], (2) a lack of pre-trained models for the specific language [17].

4. Conclusions and further work

In this work we had the opportunity to take part in the HASOC'2021 shared-task for automatic detection of hate speech on various languages. We achieved relevant results on some of the proposed tasks, with 84.23% with an ensemble model for hate content detection in Marathi and achieving a 51.67% of macro F1-score for discriminating between hate, profane and offensive content in Hindi with an ensemble model.

Although these are good results, we consider that there are improvements that can be done in order to reach better results. We propose the creation of more pre-trained models for more languages, which include Hindi and Marathi. We also consider the analysis of other models that could improve our results in these tasks. Besides, we will pay special attention to language models and linguistic features capable of extracting patterns from figurative language [18], in which words and expressions shift from their literal meaning. For this, we will compile datasets from satiric media and we will analyze linguistic devices such as sarcasm or irony, similar to the works described at [19].

Acknowledgments

This research paper is part of the research project PID2019-107652RB-I00 funded by MCIN/ AEI/10.13039/501100011033. In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme.

References

[1] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech, in: FIRE: 2021 Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.

[2] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021. URL: http://ceur-ws.org/.

[3] S. Gaikwad, T. Ranasinghe, M. Zampieri, C. M. Homan, Cross-lingual offensive language identification for low resource languages: The case of marathi, in: Proceedings of RANLP, 2021.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[5] J. A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america, Future Generation Computer Systems 112 (2020) 614–657. doi:10.1016/j.future.2020.06.019.

[6] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, Future Generation Computer Systems 114 (2021) 506 – 518. URL: http://www.sciencedirect.com/science/article/pii/S0167739X20301928. doi:10.1016/j.future.2020.08.032.

[7] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, Proceedings of the International AAAI Conference on Web and Social Media 11 (2017) 512–515. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14955.

[8] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PLOS ONE 14 (2019) 1–16. URL: https://doi.org/10.1371/journal.pone.0221152. doi:10.1371/journal.pone.0221152.

[9] T. Ranasinghe, M. Zampieri, Multilingual offensive language identification with cross-lingual embeddings, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 5838–5844. URL: https://aclanthology.org/2020.emnlp-main.470. doi:10.18653/v1/2020.emnlp-main.470.

[10] I. Bigoulaeva, V. Hangya, A. Fraser, Cross-lingual transfer learning for hate speech detection, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Kyiv, 2021, pp. 15–25. URL: https://aclanthology.org/2021.ltedi-1.3.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019). URL: http://arxiv.org/abs/1910.03771. arXiv:1910.03771.

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

[13] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[14] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[15] S. Wu, M. Dredze, Are all languages created equal in multilingual bert?, CoRR abs/2005.09093 (2020). URL: https://arxiv.org/abs/2005.09093. arXiv:2005.09093.

[16] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, CoRR abs/1906.01502 (2019). URL: http://arxiv.org/abs/1906.01502. arXiv:1906.01502.

[17] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for finnish, CoRR abs/1912.07076 (2019). URL: http://arxiv.org/abs/1912.07076. arXiv:1912.07076.

[18] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of english literature on figurative language applied to social networks, Knowl. Inf. Syst. 62 (2020) 2105–2137. URL: https://doi.org/10.1007/s10115-019-01425-3. doi:10.1007/s10115-019-01425-3.

[19] M. del Pilar Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, G. Alor-Hernández, Automatic detection of satire in twitter: A psycholinguistic-based approach, Knowl. Based Syst. 128 (2017) 20–33. URL: https://doi.org/10.1016/j.knosys.2017.04.009. doi:10.1016/j.knosys.2017.04.009.