SSN_NLP_MLRG@HASOC-FIRE2020: Multilingual Hate Speech and Offensive Content Detection in Indo-European Languages using ALBERT

A. Kalaivani[a], D. Thenmozhi[a]

[a]*Department of CSE, SSN College of Engineering, OMR, Kalavakkam, Tamil Nadu 603110*

Abstract

This paper presents our system submitted the runs to HASOC 2020: Hate Speech and Offensive Content Identification in Indo-European Languages. The detection of hate speech and offensive content in social media has much attention in recent studies. Moreover, the identification of hate speech content in various languages moves forward in the field of Natural language processing. We have participated in task1, task2 for the English, German, and Hindi (code-mixed) languages. We have adapted and fine-tuned the pre-trained ALBERT models for all the three languages. We also employed the ULMFiT framework to categorize the hate and offensive comments. We use cross-lingual translation to enrich the training data for Hindi and German languages. Our team achieved the macro-averaged F1-scores 0.88, 0.76, 0.41 in task1 for the English, German, Hindi, and 0.53, 0.50, 0.30 in task2 for the English, German, Hindi language. The final leaderboard decided to calculate the results by using 15% of private data. Our team obtained the macro-averaged F1-scores 0.4979, 0.5025, and 0.3971 in task1 for the English, German, Hindi, and 0.2305, 0.2920, 0.2063 in task2 for the English, German, Hindi language. Our team achieved the 2^{nd} rank on the private leaderboard test data in task2 for the German language.

Keywords

Hate speech detection, Offensive language detection, Cross-lingual translation, Transformers, Language modeling

1. Introduction

Nowadays, there is an immense growth of online user content in a social network such as Twitter, Facebook, YouTube, Instagram, etc. This socialization affected the life of the people, and their behavior could seriously lead to emotionally ill likes suicide, depression, frustration [1, 2]. Hate speech [1] defines the attacks against an individual or group, based on these attributes as race, gender, ethnicity, religion, sexual orientation, age, physical or mental disability, and others. Offensive content [2] such as harassment, threatening, violent, defrauding or obscene material, sexual comments, gender-specific comments, racial slurs, any content that could seriously offend someone or group based on their age, religious or political beliefs, sexual orientation, marital or parental status, physical features, national origin, or disability. So the direct hate

FIRE '20, Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India.

kalaiwind@gmail.com (A. Kalaivani); theni_d@ssn.edu.in (D. Thenmozhi)

[1]https://www.dictionary.com/browse/hate-speech

[2]https://www.lawinsider.com/dictionary/offensive-content

speech, offensive content, insults, abusive languages are identified. But the indirect comments like sarcasm [3], metaphors are challenging to annotate by humans. Therefore, Automatic detection of hate speech and offensive language is a challenging task.

The HASOC 2019 [4] competition was an aim to build systems capable of identifying hate speech and offensive content in social networks for the English, German and Hindi languages. The HASOC 2019 [3] organizers defined three sub-tasks for English, Hindi and two subtasks for German. Furthermore, whether the content is Non-Hate-Offensive or Hate and Offensive (sub-task A), what is the type of the hate and offensive content (sub-task B), and who is the target of the Hate and Offensive message (sub-task C). This year, HASOC 2020 features datasets with three languages includes English, German, and Hindi.

This article presents our approaches to HASOC-2020. We have participated in task1 and task2 for all three languages. The goal of the task1 is to recognize if a message is hate speech and offensive or not. The task2 aims to categorize the content into hate, offensive, and Profanity. We adapt and fine-tune ALBERT pre-trained model with ktrain library for all languages. We use the cross-lingual translation for German and Hindi to high resource language. For the Hindi language, we employed the ULMFiT (Universal Language Encoder Fine-Tuning for Text Classification) with the FastAi library. We used the NLTK library for pre-processing the train and test data for all languages. The structure of the paper as follows. Section 2 reviews the related works. Section 3 presents the data description and methodology of our models. We will discuss the experimental and submitted results in Section 4. Finally, Section 5 involves the conclusion of our work and discuss further improvement.

2. Related Work

The transfer learning enabled by BERT pre-trained language models, GPT, and ULMFiT, researchers have adapted to using these methods for focusing the Offensive Language detection task. In the OffensEval 2019 competition [5], the top seven teams utilized BERT with different variations in the parameter settings and the pre-processing steps. The bi-directional LSTM with various attention mechanisms and machine learning approaches used to detect the offensive language in SemEval 2019 shared task [6]. OffensEval 2020: Shared Task on Multilingual Offensive Language Identification for the English, Greek, Danish, Turkish, and Arabic languages [7], the majority of teams used pre-trained embedding's such as con-textualized Transformers, ELMo embeddings, transformers were BERT [8], RoBERTa, and the multilingual mBERT. The researchers [9] present a cross-lingual data augmentation technique by replacing a segment of the input context with its translation in another language, which is high resource languages, for example: English language.

Gemeval2018 [10] is the shared task of identification of offensive language, and they mainly focused on the offensive content in German language micro-posts. HatEval [11]: the shared task of multilingual detection of hate speech against immigrants and women in Twitter forums, and the best team enhanced an SVM model with RBF kernel and make use of sentence embedding from Google Universal Sentence Encoder. From the reviews, most of the methods are machine learning, deep learning, pre-trained transformer models that can obtain the optimal solutions.

[3] https://hasocfire.github.io/hasoc/2019/dataset.html

Table 1
Annonated Tweets - HASOC2020

Tweets	Task1	Task2
I fucking love life	**HOF**	**PRFN**
@viddywel2 Think you're funny	NOT	NONE
Look at that ass!!	HOF	OFFN
@BeaurisZ Is ja auch fürn Arsch	HOF	PRFN
Geh zurück ins Loch.DUMMES Individuum.	**HOF**	**OFFN**
put yo face in his ass	**HOF**	**HATE**

We still face the problems of handling the imbalanced dataset in various languages and annotated the data by humans, mainly code-mixed languages. This problem leads to open research in different languages other than English. HASOC 2020 shared task provides the resource for the English, German, and Hindi (code-mixed) languages.

3. Data and Methodology

3.1. Data Description

Typically the HASOC 2020 dataset [4] offers posts from Twitter, YouTube, and Facebook. The training dataset is given in the .xlsx file. The test data is distributed in a comma-separated format. The dataset includes posts written in English, German, and Hindi (code-mixed) languages. For the English language, the training data size is approximately 3708 posts and the size of the test data is 814 posts. For the German language, the size of the training data is 2373 posts and the test data size is 526 posts. For the Hindi language, the training data size is 2963 and the test data size is approximately 663 posts. Table 1 presents the annotated tweets for the English, German, Hindi languages from the HASOC 2020 dataset.

In HASOC 2020, each language includes two tasks, namely task1 and task2. Task1 is a binary classification task in which the aim is to build systems able to classify the given tweets into two classes, namely, Hate and Offensive content (HOF): the posts contain the profane, insults, threatening words. Non-Hate and Offensive (NOT): The tweets do not include hate and offensive content. Task2 is a multi-label text classification in which the aim to develop systems able to classify the tweets into three classes, namely (HATE) Hate speech: the posts which contain hate words. (OFFN) Offensive: the post which contains offensive content. (PRFN) Profane: the post contains profanity words. We participate in task1 and task2 for all the three languages.

3.2. Data preprocessing

The data preprocessing were minimal to make that portable for all the task of the particular languages. We perform a specified task by using NLTK [5] libraries for this data in the three languages. First, we remove the strings start with @ symbol because string denoted as the name

[4]https://hasocfire.github.io/hasoc/2020/dataset.html
[5]https://www.nltk.org/

Table 2
Validation scores of ALBERT model for the three Languages

Task	Accuracy	Macro F1	Weighted F1
English Task1	0.91	0.91	0.91
English Task2	0.81	0.50	0.77
German Task1	0.84	0.81	0.84
German Task2	0.80	0.40	0.75
Hindi Task1	0.72	0.42	0.60
Hindi Task2	0.72	0.21	0.60

of the users, and it does not contain expressions and also reduces the performance of the model. After that, we remove the hashtags, punctuations, URLs, numerals because usually, the string starts with https://ab, which doesn't have semantic meaning. So it is considered as noisy data. We finally remove the emoji's and then convert all the upper case text into lower case text.

3.3. Methodology

We used the pre-trained models ALBERT (A Lite BERT for self-supervised learning of language representations) with the ktrain [6] library to build the systems for the English, German and Hindi language. We also used the ULMFiT framework with the FastAi [7] library to build the system, especially for the Hindi language. We know that ktrain is a lightweight wrapper in TensorFlow 2 (tf.keras), and it helps to build, train, deploy the neural networks, machine learning models, and it makes the deep learning models more accessible.

For the validation process, we take 0.2 of test data from the training data and set the random-state is to 100. We used the Albert-base-v2 model to build, train, predict the data for all the three languages. Firstly, we create an instance by using model, sequence length, target class for transformer in ktrain, and set the sequence length is 512. usually, The data should be truncated, which is greater than the sequence length. After that, we preprocess the data that is suitable for the ALBERT model. Next, we fine-tune the classifier with the pre-trained weights and randomly chooses the Final layers. We set the batch size to 6, and the model wrapped in ktrain, which is easy to train the data and predict the new data. We set the learning rate to 3e-5 and the epoch to 5. The performance of the three languages of the ALBERT model is presented in the Table 2. Separately, we translate the German and Hindi languages into English languages by using google translate. Then, we considered that cross-lingual languages as separate input for evaluation using the ALBERT model for the translated German task1 and task2 and the translated Hindi language task1 and task2.

We used the ULMFiT framework using the FastAi Library to classify the hate speech and offensive content for the Hindi language. We create the language model with AWD-LSTM (Average-SGD Weight-Dropped LSTM) architecture model for the text classification to predict the hate and offensive content. The fastai library provides functions to create classification and Language model data bunch. we set the batch size to 32, the learning rate to 3e-02, 3e-04, 3e-03,

[6]https://pypi.org/project/ktrain/
[7]https://docs.fast.ai/

Table 3
Test sample Results of English, German, and Hindi Languages

Task	Precision	Recall	Macro F1	Weighted F1
English Task1	0.887	0.886	0.884	0.884
English Task2	0.580	0.531	0.534	0.811
German Task1	0.759	0.771	0.765	0.819
German Task2	0.359	0.411	0.383	0.747
Hindi Task1	0.351	0.5	0.412	0.580
Hindi Task2	0.185	0.25	0.213	0.634
Trans_German Task1	0.774	0.742	0.755	0.819
Trans_German Task2	0.547	0.482	0.501	0.772
Trans_Hindi Task1	0.351	0.5	0.412	0.580
Trans_Hindi Task2	0.285	0.328	0.305	0.694

1e-03, 5e-03, 5e-04, and the epoch to 1, 15, 3, 2 and 5 for training. We have got a Macro-averaged F1-score of 0.4 and 0.2 with epochs 3 for task1 and task2 Hindi language. Comparatively, the ALBERT model performed well than the ULMFiT model for the Hindi language.

4. Results

In this section, we present the evaluation of our model as well as the evaluation of the final submissions. We have submitted the best model after comparing the performance of our models for English, German, and Hindi languages.

4.1. Experimental Results

We used the evaluation metrics like precision, recall, macro-averaged F1-score, and weighted-average F1-score. The HASOC 2020 organizers provide the test data for the English, German, and Hindi languages. Based on the validation performance, we fine-tune the ALBERT model to build and predict the data for all the three languages. The performance is analyzed, and the test sample results of the ALBERT model for the three languages are presented in the Table 3. Comparatively, the task1 and task2 of the English language performs better than the other languages because of the large amount of data. For the German language, the translated German language of task1 and task2 performs well than actual German language task1 and task2 because of the availability of high resources in the English language. For the Hindi language, we used the ALBERT model and ULMFiT model in that the ALBERT model performs well in translated Hindi language, especially for task2.

4.2. Submitted Results

We present the final results of the evaluation of our submissions for all the languages. The HASOC 2020 organizers [12] have carried out the evaluation process. They have chosen approximately 15% of private test data for evaluation. The task organizers provided the evaluation

Table 4
Final Results of English, German, and Hindi Languages

Task	Macro F1
English Task1	0.4979
English Task2	0.2305
German Task1	0.5025
German Task2	0.2920
Hindi Task1	0.3971
Hindi Task2	0.2063

report based on macro-averaged F1-scores. Our team SSN_NLP_MLRG submission had macro-averaged F1-scores 0.4979, 0.5025, and 0.3971 for task1 of the English, German, and Hindi, respectively. Our team submission had macro-averaged F1-scores 0.4979, 0.5025, and 0.3971 for task1 of the English, German, and Hindi, respectively. Furthermore, our team submission had macro-weighted F1-scores 0.2305, 0.2920, 0.2063 for task2 of the English, German, and Hindi, respectively. The F1-scores of English task1, German task1, Hindi task1 has improved when compared with the baseline F1-scores. Our team achieved the 2^{nd} rank on the final private leaderboard test data in task2 for the German language. The final private leaderboard results of the English, German, and Hindi language of task1 and task 2 are presented in Table 4.

5. Conclusions

In this paper, we presented the system for identifying the hate speech and offensive language in tweets, youtube comments, and Facebook posts in English, German, and Hindi languages. Our system uses minimal preprocessing techniques. We experimented with a pre-trained ALBERT transformer model with the variations of inputs and ULMFiT to determine the most suitable model for the tasks in three languages. According to our evaluation, the results provided by the task organizers, it is clear that fine-tuning ALBERT architecture scores well. Our model performs well in all three languages. Due to non-language-specific preprocessing, we used cross-lingual translation for better performance. In future research, we can improve performance by using different classification algorithms. Further, we will extend this system to other languages.

References

[1] E. Whittaker, R. M. Kowalski, Cyberbullying via social media, Journal of School Violence 14 (2015) 11–29. URL: https://doi.org/10.1080/15388220.2014.949377. doi:10.1080/15388220.2014.949377. arXiv:https://doi.org/10.1080/15388220.2014.949377.

[2] A. Kalaivani, D. Thenmozhi, Sentimental analysis using deep learning techniques, International Journal of Recent Technology and Engineering (IJRTE) 7 (2019) 600–606.

[3] A. Kalaivani, D. Thenmozhi, Sarcasm identification and detection in conversion context using BERT, in: Proceedings of the Second Workshop on Figurative Language Processing,

Association for Computational Linguistics, Online, 2020, pp. 72–76. URL: https://www.aclweb.org/anthology/2020.figlang-1.10. doi:10.18653/v1/2020.figlang-1.10.

[4] 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, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[5] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), 2019. arXiv:1903.08983.

[6] D. Thenmozhi, B. Senthil Kumar, S. Sharavanan, A. Chandrabose, SSN_NLP at SemEval-2019 task 6: Offensive language identification in social media using traditional and deep machine learning approaches, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 739–744. URL: https://www.aclweb.org/anthology/S19-2130. doi:10.18653/v1/S19-2130.

[7] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Çağrı Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), 2020. arXiv:2006.07235.

[8] A. Kalaivani, D. Thenmozhi, SSN_NLP_MLRG at SemEval-2020 task 12: Offensive language identification in English, Danish, Greek using BERT and machine learning approach, in: Proceedings of the International Workshop on Semantic Evaluation, 2020.

[9] J. Singh, B. McCann, N. S. Keskar, C. Xiong, R. Socher, XLDA: cross-lingual data augmentation for natural language inference and question answering, CoRR abs/1905.11471 (2019). URL: http://arxiv.org/abs/1905.11471. arXiv:1905.11471.

[10] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the germeval 2018 shared task on the identification of offensive language, in: In Proceedings of GermEval, 2018.

[11] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://www.aclweb.org/anthology/S19-2007. doi:10.18653/v1/S19-2007.

[12] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages), in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR, 2020.