LAs for HASOC- Learning Approaches for Hate Speech and Offensive Content Identification

F Balouchzahi, H L Shashirekha

Department of Computer Science, Mangalore University, Mangalore - 574199, India

Abstract

Anti-social elements in social media take advantage of the anonymity in the cyber world and indulge in vulgar and offensive communications such as bullying, trolling, harassment etc. Many youths experiencing such victimization are reported to have psychological symptoms of anxiety, depression and loneliness. These issues have become a growing concern for society and hence, it is important to identify and remove such behaviors in the society at the earliest. In view of this, this paper describes the learning models proposed by our team MUCS, for identifying hate speech and offensive content. Three architectures based on different learning approaches namely Ensemble of Machine Learning (ML) algorithms, Transfer Learning (TL) and ML-TL - a hybrid combination of the first two approaches are proposed. Our team obtained macro f1-score of 0.4979, 0.2517, 0.5044 and 0.5182 for English Subtask A, Subtask B, German Subtask A and Hindi Subtask A respectively.

Keywords

Learning Approaches, HASOC, Machine Learning, Transfer Learning, Ensemble, ULMFiT

1. Introduction

Social media analysis is important for many companies such as Facebook, Instagram and even online shopping websites and this analysis includes various tasks such sentiments analysis, hate speech detection, etc. Speed of spreading Hate Speech and Offensive Content (HASOC) is increasingly becoming higher due to the rapid development in mobile and web technology [1]. These contents can have negative impact on the society especially on the younger generation as they will be more active on online platforms[1]. Further, situations like covid-19 pandemic and nuclear families are creating an addiction to online platform for the younger generation. Reports of anti-social elements in social media taking advantage of the anonymity in the cyber world and targeting younger generation and women are increasing day by day. Many youths experiencing such victimization are reported to have psychological symptoms of anxiety, depression, and loneliness. These issues have become a growing concern for the society and therefore it is important to identify and remove such behaviors in the society at the earliest.

Detecting hate speech and offensive content in order the curb it's spreading at the early stage

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

frs_b@yahoo.com (F. Balouchzahi); hlsrekha@gmail.com (H.L. Shashirekha)
https://mangaloreuniversity.ac.in/dr-h-l-shashirekha (H.L. Shashirekha)
0000-0003-1937-3475 (F. Balouchzahi)

[1] https://www.internetmatters.org/hub/question/what-is-the-real-world-impact-of-online-hate-speech-on-young-people

is the need of the hour. In this direction, we, team MUCS present three Learning Approaches (LA) namely, i) Ensemble of Machine Learning (ML) algorithms using word/character n-grams features, ii) Transfer Learning (TL) using Universal Language Model Fine-Tuning ULMFiT[2] model that use a pre-trained Language Model (LM) and fine-tuning that LM for identifying hate speech and offensive contents and iii) ML-TL, a hybrid combination of the first two approaches, for the identification of hate speech and offensive content in Indo-European Languages namely, English, Germany and Hindi in shared task called Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) 2020[3] in Forum for Information Retrieval Evaluation (FIRE) 2020[4]. The HASOC involves 2 subtasks for each language: i) Subtask A is a typical binary classification problem which identifies whether a given text contains "HOF" i.e., hate, offensive and profane content or "NOT" i.e., no hate, offensive and profane content. ii) Subtask B is a multi-class classification problem of identifying whether the "HOF" labeled text in Subtask A contains hate speech, offensive or profane content and labeling it as HATE, PRFN or OFFN respectively. More details about the tasks are given in competition page and reference paper [2].

2. Related Work

HASOC is not a new challenge and till now many studies have been done in this area including HASOC 2019[5], a shared task in three languages namely, English, German, and Hindi. The organizers [3] of the shared task developed three datasets for each language collecting data from Twitter and Facebook and organized three subtasks namely, Subtask A, Subtask B and Subtask C for each language. Subtask A is a binary classification of Hate Speech (HOF) and non-offensive content. If the post in Subtask A is identified as HOF, then Subtask B is to identify the type of hate and Subtask C is to identify whether the post is targeted or not. Some of the works related to HASOC are given below:

Two studies for fake news spreader detection based on different learning approaches for English and Spanish languages have submitted to PAN 2020 shared task by Shashirekha et. al. [4][5]. Datasets were provided by PAN 2020 [6] for training the models. An ensemble voting classifier of the three classifiers (two Linear SVC classifiers and a Logistic Regression) built by Shashirekha et. al. [5] using Unigram TF/IDF, N_gram TF and Doc2Vec feature sets obtained 73.50% and 67.50% accuracies for English and Spanish languages respectively. In another work proposed by Shashirekha et. al. [4], TL model based on ULMFiT is initially trained on a general domain English/Spanish data collected from Wikipedia which is then fine-tuned using target task dataset and used for the fake news spreader detection task as the target model. Their models obtained 62% and 64% accuracies on English and Spanish languages respectively.

In the system based on ordered neurons LSTM proposed by Wang et. al. [7], they utilized an attention layer to assign a weight to each word in the sentence to reveal the words contributing to the offensive character of a post more prominently. For HASOC OLID [8] datasets their

[2] Universal Language Model Fine-Tuning

[3] https://hasocfire.github.io/hasoc/2020

[4] http://fire.irsi.res.in/fire/2020/home

[5] https://hasocfire.github.io/hasoc/2019/index.html

model achieved first rank in Subtask A for English language with the macro f1-score of 0.7882 and the weighted f1-score of 0.8395. A multilingual LSTM model that has been submitted to HASOC 2019 by Tharindu et. al. [9] obtained 3rd rank in Subtask A for English language. The authors used 7 different architectures based on neural networks namely, pooled Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM) and GRU with Attention, 2D Convolution with Pooling, GRU with Capsule and LSTM with Capsule and Attention, and a fine-tuned BERT model which achieved best results among other models with macro f1-score of 0.7891, 0.5881, 0.8025 for English, Germany, and Hindi respectively. Shubhanshuet. al. [10] submitted a fine-tuning pre-trained monolingual and multilingual transformer (BERT) using neural network models for HASOC 2019 and obtained first place for English subtasks B and C with a macro average f1-score of 0.5446 and 0.5111 respectively and obtained a macro average f1-score of 0.5812 for Hindi subtask B. They also tried a joint-label based approach called shared-task D to alleviate data sparsity in shared tasks, while achieving competitive performance in the final evaluation.

Victor et. al. [11] presented two different approaches on HASOC 2019. In the first approach they combined CNNs and RNNs for handling n-grams and long-term dependencies utilizing three embedding layers as inputs namely, embedding of a pre-processed post, embedding of its Part of Speech (POS) tagging, and the existence of positive or negative words, according to a pre-defined lexicon. Second approach is based on LSTM networks with an attention layer to focus on critical words in the sentence that takes the embedding representation of respective pre-processed post, together with features extracted from the tweet's POS tagging. These features together with GloVe word embeddings were tested on conventional ML models (SVM, LR and Naive Bayes) and some simple DL models (MLP, CNN and Simple Dense Layer). The ensemble of ML and DL models obtained 3rd rank in Subtask A and 2nd rank in Subtask B and C with weighted f1-score of 0.8182, 0.7595, and 0.7840 in subtasks A, B and C respectively.

3. Methodology

We, team MUCS carried out various experiments on different learning models and the best performing models are submitted for each task of all languages. Once binary classification for Subtask A is done dataset has been filtered with "OFF' labels (only offensive posts are used to train models for Subtask B) and then models have been trained on filtered data.

3.1. Architectures

The details of learning models used in this study are given below:

ULMFiT TL: The architecture of ULMFiT TL introduced by Howard et. al. [12] is based on the concept of transferring the knowledge gained in developing one task called source task to develop another task called target task, instead of starting the target task from scratch [12][13]. Stages of ULMFiT TL are shown in figure 1. In the proposed ULMFiT model, pre-trained LM (a probability distribution over word sequences in a language [12]) which represents the general features of a language is used as source model and the knowledge obtained from LM along with the train set is used in building a target model i.e., a hate speech detection model in this study.

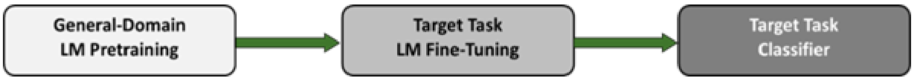

Figure 1: Stages in ULMFiT model

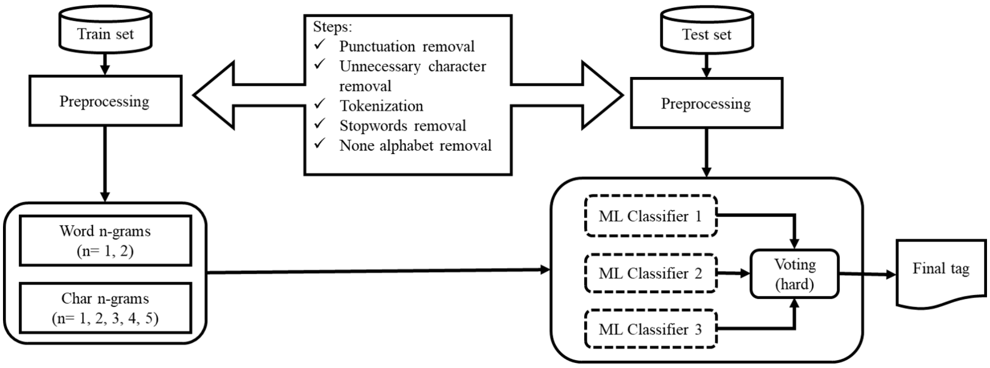

Figure 2: Ensemble of ML models

The proposed ULMFiT model utilizes an encoder for an ASGD Weight-Dropped LSTM (AWD-LSTM) that consists of a word embedding of size 400, 3 hidden layers and 1150 hidden activations per layer which can be plugged in with a decoder and classifying layers to create a text classifier [12][14]. Target classifier is created using text.models module from fastai[6] library which is based on TL.

Ensemble of ML algorithms: Ensembling ML algorithms generally mean utilizing the strength of several ML classifiers by different methods to get better results. In this work, the classifiers are trained on word/char n-grams and majority voting is used for ensembling. Figure 2 illustrates the architecture of the ensemble of ML models.

ML-TL: In this hybrid approach, instead of 3rd ML classifier shown in figure 2, ULMFiT TL model has been used and the majority voting of labels predicted by both ML and TL models is used to assign labels to the given text.

3.2. Languages and Subtasks

The models designed for the subtasks of each language are shown in Table 1. Publicly available pre-trained LM by Howard Jeremy and Sebastian Ruder [12] trained on the WikiText-103 dataset is used for ULMFiT model. Ensemble ML with 'hard' voting is used for both subtasks of German language with a maxdepth of 100 for Random Forest. Publicly available pre-trained Hindi LM[7] is used as source task in ULMFiT TL for Hindi language.

[6] https://docs.fast.ai/

[7] https://github.com/goru001/nlp-for-hindi

Table 1
Models for the subtasks of each language

Subtasks	English	German	Hindi
Subtask A	ML-TL (SVC, LR, ULMFiT)	Ensemble ML (RFC, LR, SVC)	ML-TL (Linear SVC, LR, ULMFiT)
Subtask B	ULMFiT	Ensemble ML (RFC, LR, Linear SVC)	ULMFiT

Table 2
Dataset statistics

Subtasks	No. of posts	Train set			Develop set		
		English	German	Hindi	English	German	Hindi
Subtask A	HOF	1856	673	847	423	134	197
	NOT	1852	1700	2116	391	392	466
Subtask B	PRFN	1377	387	148	293	88	27
	HATE	158	146	234	25	24	56
	OFFN	321	140	465	82	36	87
	NONE	1852	1700	2116	414	378	493
Total		3708	2373	2963	814	526	663

4. Experimental Results

4.1. Datasets

In this shared task, train set and development set for each language is provided by task organizers and after submitting code, weights (if any) and results, models were tested by the organizers on 15% of private test set. Details and statistics of datasets provided by HASOC 2020 [2] are given in Table 2. Statistics of datasets show that for Subtask A, English training and development set are balanced but German and Hindi training and development set are not balanced. But, for Subtask B English training and development set are heavily imbalanced and German and Hindi training and development set are not balanced. Further, as all posts with 'NOT' labels in Subtask A will be 'NONE' for Subtask B, these posts (posts with 'NOT' labels) are excluded during training models for Subtask B and for submission they have added with 'NONE' label directly.

4.2. Results

The results obtained on development set using sklearn.metrics[8] module for each language is shown in Table 3. ML-TL model for English language Subtask A has obtained best performance with a macro f1-score of 0.87 and ULMFiT TL model for Hindi language Subtask B has obtained the highest performance with a macro f1-score of 0.70. Performance of the models was evaluated by task organizers using 15% of the private test set and the results of proposed models

[8] https://scikit-learn.org/stable/modules/classes.htmlsklearn-metrics-metrics

Table 3
Results on development set

Subtasks	English		German		Hindi	
	Architecture	macro f1	Architecture	macro f1	Architecture	macro f1
Subtask A	ML-TL	0.87	Ensemble ML	0.79	ML-TL	0.71
Subtask B	ULMFiT TL	0.62	Ensemble ML	0.64	ULMFiT TL	0.70

Table 4
Results announced by the organizers on 15% of private test data

Subtasks		English		German		Hindi	
		Rank	macro f1	Rank	macro f1	Rank	macro f1
Subtask A	Obtained by MUCS	21	0.4979	11	0.5044	8	0.5182
	Best in HASOC 2020	1	0.5152	1	0.5235	1	0.5337
Subtask B	Obtained by MUCS	5	0.2517	XX	XX	XX	XX
	Best in HASOC 2020	1	0.2652	1	0.2943	1	0.3345

and the best performance reported in corresponding subtask are as shown in Table 4. Results reported by the organizers show a close competition among the participated teams and our best performance on Subtask A in Hindi with 0.5182 macro f1-score obtained 8th rank, and Subtask B in English with 0.2517 macro f1-score obtained 5th rank. However, in the subtasks of all languages, the proposed approaches achieved results with a difference of 0.02% to first rank.

5. Conclusion

In this paper, we describe Machine Learning and Transfer Learning approaches proposed by our team MUCS for Hate Speech and Offensive Content Identification (HASOC) shared task in FIRE 2020. The results illustrate that the proposed approaches obtained overall good and competitive results that are close to the highest reported results on each subtask. As future work, we would like to explore different learning approaches on code-mixed and native languages.

6. Acknowledgments

We would like to thank and appreciate team HASOC 2020 organizers for all valuable supports and comments during this task.

References

[1] Z. Mossie, J.-H. Wang, Vulnerable community identification using hate speech detection on social media, Information Processing & Management 57 (2020) 102087.

[2] 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). URL: http://ceur-ws.org/.

[3] 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, (2019), pp. 14–17.

[4] F. Balouchzahi, H. L. Shashirekha, Ulmfit for twitter fake news spreader profiling - notebook for PAN at CLEF 2020, In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol, editors, CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org (2020).

[5] M. D. Anusha, H. L. Shashirekha, N. S. Prakash, Ensemble model for profiling fake news spreaders on twitter - notebook for PAN at CLEF 2020, In Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol, editors, CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org (2020).

[6] 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).

[7] B. Wang, Y. Ding, S. Liu, X. Zhou, Ynu_wb at hasoc 2019: Ordered neurons lstm with attention for identifying hate speech and offensive language., in: FIRE (Working Notes), (2019), pp. 191–198.

[8] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, A hierarchical annotation of offensive posts in social media: The offensive language identification dataset, arxiv preprint (2019).

[9] T. Ranasinghe, M. Zampieri, H. Hettiarachchi, Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification., in: FIRE (Working Notes), (2019), pp. 199–207.

[10] S. Mishra, S. Mishra, 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages., in: FIRE (Working Notes), (2019), pp. 208–213.

[11] V. Nina-Alcocer, Vito at hasoc 2019: Detecting hate speech and offensive content through ensembles., in: FIRE (Working Notes), (2019), pp. 214–220.

[12] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, arXiv preprint arXiv:1801.06146 (2018).

[13] S. Faltl, M. Schimpke, C. Hackober, Ulmfit: State-of-the-art in text analysis (2019).

[14] S. Merity, N. S. Keskar, R. Socher, Regularizing and optimizing lstm language models, arXiv preprint arXiv:1708.02182 (2017).