Application of XLM-RoBERTa for Multi-Class Classification of Conversational Hate Speech

Tebo Leburu-Dingalo, Karabo Johannes Ntwaagae, Nkwebi Peace Motlogelwa, Edwin Thuma and Monkgogi Mudongo

Department of Computer Science, University of Botswana

Abstract

In this paper, team University of Botswana Computer Science (UB-CS) investigate the use of XLM-RoBERTa, a multilingual model trained on 100 different languages for transfer learning in the identification of conversational hate-speech in code-mixed languages. We also investigate whether enriching the tweets with textual sentiments from emojis can help improve the classification performance. Our proposed solution outperformed other teams that participated at the HASOC (2022) Task 2 with a macro F1 score of 0.4939. The result suggest that enriching the tweets with textual sentiments and using a pre-trained multilingual model for transfer learning can help in the identification of conversational hate-speech in code-mixed languages.

Keywords

Hate Speech, XLM-RoBERTa, Transfer Learning

1. Introduction

Social Media sites provide opportunities for users regardless of background to post their view about different topics. Inadequate restrictions on these sites has however led to the wide publication of inappropriate content on the web. This includes the posting of fabricated comments and use of hateful, abusive and offensive language often intended to degrade or cause harm to the image or status of other individuals or certain groups in the society [1]. Such content has also on occasion been thought to lead to gruesome events such as unprovoked attacks on individuals or groups. In an effort to curb this practice there has been a growing effort in recent years to develop automated systems to monitor user posts for hateful or offensive content on social networking platforms [2]. According to Modha et al. [2], a notable shortcoming with many of these systems is that they treat each post as an individual entity and thus do not pay attention to contextual elements within which the post was made. This is problematic since most of the messages in social media form part of a conversational thread, meaning a seemingly inoffensive message might actually be offensive considering the overall context of the thread. For instance a positive message might be judged as not offensive while in fact

Forum for Information Retrieval Evaluation, December 9-13, 2022, India

leburut@ub.ac.bw (T. Leburu-Dingalo); ntwaagaek@ub.ac.bw (K. J. Ntwaagae); motlogel@ub.ac.ub.bw (N. P. Motlogelwa); thumae@ub.ac.bw (E. Thuma); mudongom@ub.ac.bw (M. Mudongo)

https://www.ub.bw/connect/staff/202 (T. Leburu-Dingalo); https://www.ub.bw/connect/staff/830 (N. P. Motlogelwa); https://www.ub.bw/connect/staff/1966 (E. Thuma); https://www.ub.bw/user/11462 (M. Mudongo)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3395/T7-12.pdf

it is supporting an offensive preceding or parent message. Furthermore messages are often expressed using a mix of languages, a property that needs to be factored in the development of hate and offensive content detection systems [2]. Hence towards addressing this challenge the HASOC 2022 Task 2: Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) - Multiclass Classification encourages the development of systems capable of detecting offensive or hateful content in tweets looking at the context of the parent content 1 [3, 4]. In particular systems should be able to identify those posts that are hateful or offensive as well as those that support the dissemination of hateful and offensive content. In this paper we attempt to address the problem through the use of a transformer model XLM-Roberta [5] which has been proved effective in multilingual text classification tasks. We fine-tune the model on the provided dataset. In an attempt to improve the model performance for the task, we focus on enhancing the tweets through data cleaning and text augmentation. To this end we pre-process the tweets and convert emojis which make a sizeable part of the tweets to text. Our approach based on the intuition that emojis can express the actual emotion felt by the user when typing a posts regardless of rhetoric expressed in the tweet. Therefore, we theorize that augmenting tweets with emoji descriptions will enhance model performance as they give a better reflection of sentiment and type of language used in the tweet.

2. Evolution of the HASOC Shared Task

The Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC (2019)) [1] shared task started in 2019 [6] inspired by two evaluation forums, OffensEval [2] [7] and GermanEval [3] [8]. In particular, the objective of the HASOC task was to develop data, hate speech detection technology and evaluation resources for several Indo-European languages. For example, the HASOC (2019) shared task offered 3 tasks. The first task (Sub-task A) offered in three languages (English, German and Hindi) was a binary classification task in which participants were required to classify tweet into Hate and Offensive (HOF) and Non- Hate and offensive (NOT) classes [6]. In Sub-task B, the classes in Sub-task A were further classified into three classes namely: (HATE) Hate speech, (OFFN) Offenive and (PRFN) Profane. In Sub-task C, only posts labelled as HOF were included and participants were required to check the type of offence[6]. The two types of offences were Targeted Insult (TIN) and Untargeted (UNT). HASOC (2020) Shared task did not differ that much from the preceding year (HASOC (2019)). In particular, the Sub-tasks A & B were made multilingual by joining the English, German and Hindi datasets in order to promote research on multilingual techniques [9].

A new task was introduced in HASOC (2021) [2] and HASOC (2022) [4] where participants were required to identify from a conversational thread whether a parent tweet, reply where either a standalone Hate (SHOF), Contextual Hate (CHOF) and Non-Hate (NONE). This was motivated by the fact that a majority messages on social networking sites form part of a conversational thread. Such conversational threads can contain hate and offensive content which may not be

[1] https://hasocfire.github.io/hasoc/2019/call_for_participation.html
[2] https://competitions.codalab.org/competitions/20011
[3] https://projects.fzai.h-da.de/iggsa/
[4] https://hasocfire.github.io/hasoc/2022/index.html

visible from a single comment or reply but can be determined if parent content is considered. The aim of the task is thus to detect posts that are hateful or offensive on their own, and those that support hate or offensive content of their parent posts. Hence the task defines three classes for the identification of hate and offensive language in posts as follows:

- (SHOF) Standalone Hate - This tweet, comment, or reply contains Hate, offensive, and profane content in itself.
- (CHOF) Contextual Hate - Comment or reply is supporting the hate, offence and profanity expressed in its parent. This includes affirming the hate with positive sentiment and having apparent hate.
- (NONE) Non-Hate - This tweet, comment, or reply does not contain Hate, offensive, and profane content in itself.

3. Experimental Setup

3.1. Training and Validation Dataset

The dataset for Task 2: Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) comprises twitter postings, comments and replies to each comment based on controversial stories from different topics including Temple-Mosque Controversy, Taliban and Covid Controversy. The tweets use mix of both the English and Hindi languages referred to as Hinglish. The statistics of the dataset is shown Table 1. This data was randomly split into 80% training data and 20% validation data.

Table 1
Training Dataset

Class Label	Training Tweets
SHOF	1636
CHOF	888
NONE	2390
Total	4914

3.2. Pre - Processing

The tweets were first concatenated to create conversational threads comprising parent tweets and comments as well as parent tweets, comments and replies where available. A manual exploration of the training data indicated that the tweets contained a lot of special characters, urls and emojis. We perform data cleaning by removing urls, stopwords, extra spaces and newlines. We however retain emojis which we expand to text using the emoji library[5] to augment the tweets.

[5]https://pypi.org/project/emoji

3.3. Selection of Model Parameters

In our emperical investigation we deploy a SimpleTransformers[6] Library by HuggingFace[7], which has implementation of task-specific SimpleTransformer models. In particular, we use a classification model called **ClassificationModel**, which uses a pre-trained model for the task of binary and multi-class classification. The model used is based on the HugginFace implementation of XLM-RoBERTa, a transformer based multilingual model pre-trained on CommonCrawl data containing 100 languages. XLM-RoBERTa is based on the BERT architecture and has a total of 12 layers for learning different semantic information with a classification layer built on top. Since we consider the influence of emojis in our experiments we first deployed the model with emojis omitted from the tweets using a learning rate of 1e-5 at 3 and 5 epochs respectively. We further experimented with augmented tweets similarly at a learning rate of 1e-5 at 3 and 5 epochs. All models used the AdamW optimizer. Base on the result in Table 2, we chose to use the parameter used in **Run 4 enhanced tweet** for our run submission the Task 2: Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL).

Table 2
Task 2 - Experimental Results

Experimental Runs	Learning Rate	Training Epochs	Accuracy	Macro-F1
Run 1 emojis omitted	1e-5	3	0.60	0.58
Run 2 with emojis omitted	1e-5	5	0.62	0.61
Run 3 with enhanced tweets	1e-5	3	0.61	0.58
Run 4 enhanced tweets	1e-5	5	0.63	0.62

Table 3
Task 2 - ICHCL Multiclass Leaderboard

RANK	TEAM	MACRO F1	MACRO PRECISION
1	UB-CS†	0.4939	0.5211
2	HASOC	0.4899	0.4884
3	fosu-nlp	0.4769	0.5042
4	bouchekif	0.4665	0.5315
5	hate-busters	0.4651	0.5348
6	nlplab_isi	0.4448	0.5247
7	irlab@iitbhu	0.4390	0.5534
8	ml_ai_iiitranchi	0.4164	0.4366
9	citk_isi	0.3952	0.4699
10	gunjan	0.2865	0.2764
11	sakshi hasoc	0.2548	0.2440

[6]https://github.com/ThilinaRajapakse/simpletransformers
[7]https://huggingface.co/xlm-roberta-base

4. Results and Analysis

Table 3 shows the leaderboard of the HASOC (2022) Task 2: Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL). Our team UB-CS denoted by † managed to outperform other teams. The results suggest that using multilingual model trained on several languages can improve the identification of conversational hate speech in code mixed languages (HINGLISH - Hindi-English). In addition, the results suggest that we can further improve the performance by enriching the tweets with textual sentiments generated from emojis.

5. Discussion and Conclusion

The results of our investigation suggests that enriching the tweets with textual sentiments and using a pre-trained multilingual model for transfer learning can help in the identification of conversational hate-speech in code-mixed languages. A natural progression of this work is to analyse whether a state-of-the-art performance can be attained by using an ensemble from several pre-trained multilingual models for transfer learning.

References

[1] C. O'Regan, Hate Speech Online: an (Intractable) Contemporary Challenge?, Current Legal Problems 71 (2018) 403–429. URL: https://doi.org/10.1093/clp/cuy012.

[2] 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: Forum for Information Retrieval Evaluation, FIRE 2021, Association for Computing Machinery, New York, NY, USA, 2021, p. 1–3. URL: https://doi.org/10.1145/3503162.3503176.

[3] S. Satapara, P. Majumder, T. Mandl, S. Modha, H. Madhu, T. Ranasinghe, M. Zampieri, K. North, D. Premasiri, Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages, in: FIRE 2022: Forum for Information Retrieval Evaluation, Virtual Event, 9th-13th December 2022, ACM, 2022.

[4] S. Modha, T. Mandl, P. Majumder, S. Satapara, T. Patel, H. Madhu, Overview of the hasoc subtrack at fire 2022: Identification of conversational hate-speech in hindi-english code-mixed and german language, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.

[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[6] 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.

[7] 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), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86. URL: https://aclanthology.org/S19-2010. doi:10.18653/v1/S19-2010.

[8] M. Wiegand, M. Siegel, Overview of the germeval 2018 shared task on the identification of offensive language, 2018.

[9] 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, FIRE 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.