Coarse and Fine-Grained Conversational Hate Speech and Offensive Content Identification in Code-Mixed Languages using Fine-Tuned Multilingual Embedding

Supriya Chanda[1], Sacchit D Sheth[1] and Sukomal Pal[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi, INDIA, 221005

Abstract
We are seeing an increase in hateful and offensive tweets and comments on social media platforms like Facebook and Twitter, impacting our social lives. Because of this, there is an increasing need to identify online postings that can violate accepted norms. For resource-rich languages like English, the challenge of identifying hateful and offensive posts has been well investigated. However, it remains unexplored for languages with limited resources like Marathi. Code-mixing frequently occurs in the social media sphere. Therefore identification of conversational hate and offensive posts and comments in Code-Mixed languages is also challenging and unexplored. In three different objectives of the HASOC 2022 shared task, we proposed approaches for recognizing offensive language on Twitter in Marathi and two code-mixed languages (i.e., Hinglish and German). Some tasks can be expressed as binary classification (also known as coarse-grained, which entails categorizing hate and offensive tweets as either present or absent). At the same time, others can be expressed as multi-class classification (also known as fine-grained, where we must further categorize hate and offensive tweets as Standalone Hate or Contextual Hate). We concatenate the parent-comment-reply data set to create a dataset with additional context. We use the multilingual bidirectional encoder representations of the transformer (mBERT), which has been pre-trained to acquire the contextual representations of tweets. We have carried out several trials using various pre-processing methods and pre-trained models. Finally, the highest-scoring models were used for our submissions in the competition, which ranked our team (irlab@iitbhu) second out of 14, seventh out of 11, sixth out of 10, fourth out of 7, and fifth out of six for the ICHCL task 1, ICHCL task 2, Marathi subtask 3A, subtask 3B and subtask 3C respectively.

Keywords
Hate Speech, Offensive Language, Social Media, Marathi, Code-Mixed, Multilingual BERT, GermanBERT, Hinglish,

1. Introduction

Over the last several years, the number of people using social media platforms and online forums has skyrocketed. Every day, around 500 million tweets are sent[1]. Unfortunately, the boom in social media usage has also resulted in an increase in hate speech and cyberbullying. Despite social media's various applications, there is one drawback, those with malicious intent view it as a chance to spread harsh thoughts to a larger audience. As a result, we must take

FIRE'22: Forum for Information Retrieval Evaluation, December 9-13, 2022, India
supriyachanda.rs.cse18@itbhu.ac.in (S. Chanda); sacchit.dsheth.cd.cse19@itbhu.ac.in (S. D. Sheth); spal.cse@itbhu.ac.in (S. Pal)
0000-0002-6344-8772 (S. Chanda); 0000-0002-9421-8566 (S. Pal)

large-scale measures to combat this bad information.

People today use social media to express their thoughts on a variety of issues. It allows people to express themselves and see the world through new eyes. This, however, empowers people to communicate whatever they choose, even if it is disrespectful or damaging to others. Social media's rapid expansion has transformed communication and content creation. Most young people use it for news consumption and social connection.

Hate speech is a menace to social culture and peace. Similarly, offensive speech may result in communicative radicalization. Automatic suppression measures are required to safeguard persons of all ages from being exposed to hate speech. Manual moderation is not always accurate, and there is a steady flow of material entering social media. Because hate speech has a detrimental influence on public opinion, several platforms, like YouTube, Facebook, and Twitter, have rules and procedures to filter hate speech content and other harmful behavior. This is an attempt to mitigate the negative impact that hate speech may have on society.[2]

Machine learning and deep learning algorithms have grown significantly for applications involving natural language interpretation as computer capabilities have advanced. These tools have a vast potential for detecting and removing malicious content from social media.

We are interested in detecting hate speech in tweets in this endeavor. This paper primarily reports on experiments with HASOC 2022 data. We assess several deep learning approaches, particularly multilingual models. We experimented with numerous fine-tuning strategies to see how they assist the model in categorizing tweets into a coarse and fine grain.

1.1. HASOC Tasks

The goal of HASOC 2022[3] [1] was to establish a testbed for the automated detection of hate speech and objectionable material in social media posts. Some tasks can be represented in binary classification- categorising hate and offensive tweets as either present or absent (also known as coarse-grained). Others can be expressed as multi-class classification- further classification of hate and offensive tweets into Standalone and Contextual hate (also known as fine-grained).

Task 1: ICHCL HINGLISH and GERMAN Codemix Binary Classification [4].

- **(NOT) Non Hate-Offensive -** Any form of Hate speech, profane, offensive content is not present in this post.
- **(HOF) Hate and Offensive -** Contains Hate, offensive, and profane words.

Task 2: Identification of Conversational Hate-Speech in Code-Mixed Languages (ICHCL) - Multiclass Classification.

- **(SHOF) Standalone Hate -** Contains Hate, offensive, and profane words in itself.
- **(CHOF) Contextual Hate -** Comment or reply of a tweet supports the hate, offense, and profanity expressed in its parent tweet. This includes expressing apparent hatred and endorsing the hatred with positive sentiment.

[1] https://hasocfire.github.io/hasoc/2022/call_for_participation.html

- **(NONE) Non-Hate -** Any form of Hate speech, profane, offensive content is not present.

Task 3: Offensive Language Identification in Marathi- focused on hate speech and offensive language identification is offered for Marathi [5, 6].

A) Offensive Language Detection-

- **Offensive(OFF)-** Contains any form of non-acceptable language
- **Non Hate-Offensive(NOT)-** No offense or profanity is present.

B) Categorisation of Offensive Language-

- **Targeted Insult (TIN)-** An insult or threat to an individual, group, or others.
- **Untargeted (UNT)-** Profanity and targeting that is untargeted.

C) Offense Target Identification-

- **Individual (IND)-** Posts targeting an individual.
- **Group (GRP)-** Posts targeting a group of people.
- **Other (OTH)-** This target is neither an individual nor a group of people.

Table 1 provides examples of the various posts and associated labels.

2. Related Work

Automated hate and offensive speech detection on social media platforms is a difficult task, for which many different machine learning and deep learning approaches have been tested. These include models trained on curated datasets, as well as models that are trained on a corpus of malicious content.

When using machine learning with hate text, there are several ways that may be done. Feature extraction is a common technique. Linear Support Vector Machines trained on TF-IDF feature model has been used[7]. This procedure may incorporate a collection of words, n-grams, lexical characteristics, and linguistic features. Word embedding algorithms have recently been proposed for similar purposes. Using the bag of words approach may result in a large number of false positives since objectionable terms in a non-hate tweet may be misclassified as hate speech[8]. [9] proposes use of pre-trained word embeddings and max/mean pooling from basic, fully connected embedding transformations was proposed as a neural-network-based hate speech categorization solution. However, these techniques fall short of capturing the whole context of the speech.

Deep learning algorithms are becoming increasingly popular in text categorization, sentiment analysis, language modelling, machine translation, and other fields.Some of these methods are Convolutional Neural Networks(CNNs)[10] [11], Recurrent Neural Networks(RNNs) [12] [13], Long Short-Term Memory(LSTMs) [14], Bidirectional LSTMs (BiLSTMs) [15] and the most

Table 1
Example tweets from the HASOC2022 dataset for all classes

Language	Sample tweet from the class	Task 1	Task 2	
Hinglish	@Joydas @NSaina 2 rs k liye tweet kiya isne samjho bhai national hero hogi phle ab to izzat gawa di.. Andhbhakt ban gayi didi	HOF	CHOF	
	@itsoutrageeyash @NSaina Yet another man telling a woman what to do and yet another telling her to #shutup. Shame on you Yash! #Feminism #Mansplaining #shethepeople	HOF	SHOF	
	No nation can claim itself to be safe if the security of its own PM gets compromised. I condemn, in the strongest words possible, the cowardly attack on PM Modi by anarchists.#BharatStandsWithModi #PMModi NOT NONE	NOT	NONE	
German	@M_Ziesmann Der Vergleich mit Hannibal Lecter alleine macht dich zur lebenden Legende!!! Ich lache tr00e4nen :D recht du damit einfach hast! Dieses ẄesenL̈auterbach, geh00f6rt einfach f00fcr immer in eine Gummizelle gesperrt mit 18h Zwangsjacke pro Tag	HOF		
	@Superutz @hartmann_torben @PrienKarin @Karl_Lauterbach @OlafScholz @_FriedrichMerz @RobertHabeckMdB	NONE		
Language	**Sample tweet from the class**	**Task 3**		
		A	**B**	**C**
Marathi	तू रेगे मध्ये येडाच होतास बिगबॉस मध्ये मंद आणि नवीन सीरियल मध्ये चोम्या जाऊदे तू टेप लाव की थोबाड शिव तीच असणार आहे	OFF	TIN	IND
	शिवसेनेने निवडणुकीत स्वतंत्र बाणा ठेवावादळभद्री भाजपबरोबर मैत्री अजिबात नको	OFF	TIN	GRP
	नेता दुई प्रजातिका हुन्छन् १ चुतिया२ कुतिया	NOT	-	-

recent is a transformer-based architecture namely Bidirectional Encoder Representations(BERT) [16][17] and XLM-Roberta[2].

Because of the scarcity of relevant corpora, the vast majority of studies on abusive language have focused on English data. To address this there have been studies on other languages like Spanish[18],French[19], Italian[20], German[21] among others.

3. Dataset

The dataset is in the form conversational thread can also contain hate and offensive content it which is not apparent just from the single comment or the reply to comment but can be identified if given the context of the parent content. The figure 1 shows the structure of data. The corpus collection and class distribution is shown in Table 2.

Figure 1: Structure of the data.

4. Methodology

4.1. Preprocessing

Twitter data are very unstructured and include a lot of noise due to the colloquial character of Twitter conversations, which might compromise accuracy of processing techniques. As a result, it was determined that all Tweets should be preprocessed to eliminate less predictive text elements. To produce the final text sequence, we concatenate the tweet and its comments and responses, if any are present. Our assumption is that this concatenation will help the model better comprehend the context, particularly in circumstances when the remark or reply is not hateful but demonstrates support for the terrible parent tweet.

- We perform cleaning by removing usernames, punctuation and URLs.
- We use ekphrasis which is a text processing tool, geared towards text from social networks, such as Twitter or Facebook. ekphrasis performs tokenization, word normalization, word segmentation (for splitting hashtags) and spell correction, using word statistics from 2 big corpora[22]
- demoji to accurately remove and replace emojis in text strings.

Table 2
Statistical overview of the Training Data and Test Data for determining the final results

			Task-1		
Data	**Language**	**# of sentences**	**NOT**	**HOF**	
Train	German Hinglish	307 4914	219 2390	88 2524	
Test	German Hinglish	81 996			
			Task-2		
			NONE	SHOF	CHOF
Train	Hinglish	4914	2390	1636	888
Test	Hinglish	996			
			Task-3-Subtask-A		
			NOT	OFF	
Train	Marathi	3103	2034	1069	
Test	Marathi	510			
			Task-3-Subtask-B		
			TIN	UNT	
Train	Marathi	1068	741	327	
Test	Marathi				
			Task-3-Subtask-C		
			IND	GRP	OTH
Train	Marathi	740	503	157	80
Test	Marathi				

4.2. Implementation

Text classification, in which we must label the dataset, is one of the problems in this data challenge. Using the given dataset, the models were developed by fine-tuning a pre-trained language model. We chose German BERT as our pretrained language model due to its recent success, as it allows working with text data in German to be more efficient with their natural language processing (NLP) tasks, and XLM-Roberta, which is capable of processing text from 100 separate languages and is trained on significantly more training data than BERT.[23]

In order to classify text more accurately, we concatenate the data in a way that provides more context to the transformer. The figure 2 shows the used concatenation process.

We have used pre-trained transformer models from HuggingFace [2] in the implementation. The Framework for Adapting Representation Models, or FARM, is based on transformers and includes extra capabilities to make developers' lives easier. Parallelized preprocessing, highly modular architecture, multi-task learning, experiment tracking, simple debugging, and close connection with AWS SageMaker are among the features.

[2]https://huggingface.co/

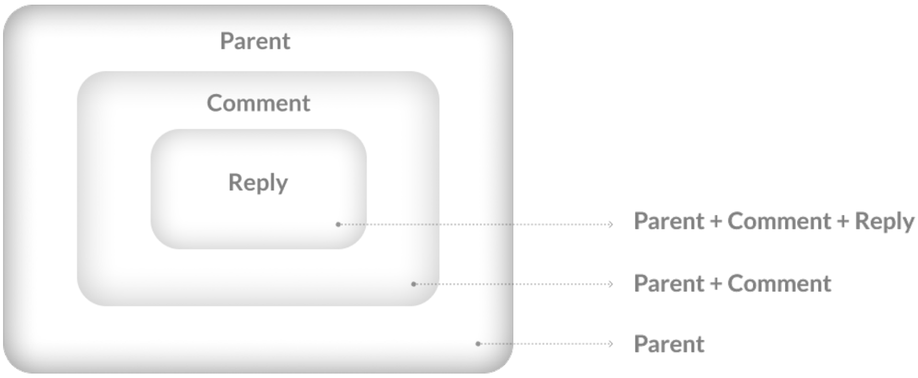

Figure 2: Concatenating text data

We use the GermanBERT [3] model, which has been fine-tuned to our standards. Despite the fact that there are numerous models trained on German data that are accessible from HuggingFace, we believe that the original model will struggle to grasp the smaller chunks of words when they are broken up and have thus employed German BERT. We utilised the PyTorch package for our implementation environment, which supports GPU processing. We discovered that training our classifier with a batch size of 16 for 5 to 10 epochs and the AdamW optimizer with a learning rate of 2e-5 worked well by doing tests. In our work, we have used the transformer-based XLM-RoBERTa model.

- Perform preprocessing for the concatenated text according to the steps mentioned in Section 4.1 .
- Individual entries are indicated with a black dot, a so-called bullet.
- We then proceed by tokenizing the text with the XLM-RoBERTa pre-trained SentencePiece tokenizer [24].
- The text is then padded and truncated to a maximum sequence length of tokens.
- The model is fine-tuned for hate speech detection with various batch sizes and AdamW optimizer.

5. Results and Discussion

Task-1 of HASOC 2022 was an ICHCL- Binary Classification problem. We provided two submissions out of which the best F_1 score we got is 0.627. We finished sixth (previously second with F_1 score 0.702) out of fourteen teams, and our submission was second out of all the runs (42) submitted by the teams. GermanBERT scores well in Task 1 for monolingual text categorization

[3] https://www.deepset.ai/german-bert

(for German), with an F1 score of 0.702. Despite the fact that several multilingual models are trained on the same dataset, a specific monolingual model can better comprehend the context of the supplied language without clutter.

Task 2 was an ICHCL- Multiclass Classification problem. We provided five submissions, of which the best F_1 score we got was 0.439. We finished seventh out of eleven teams, and our submission was tenth out of all the runs (26) submitted by the teams. mBERT scores well in task 2 which was a fine-graind conversational Hate speech and offensive content identification.

In Task 3, which was offensive language identification in Marathi, had three subtasks. For subtask-A, we provided five submissions, out of which the best F_1 score we got is 0.935. We finished sixth out of the ten teams, and our submission was ninth out of all the runs (28) submitted by the teams. We submitted one run for each Subtask-B Subtask-C, resulting in an F_1 score of 0.535 and 0.289, respectively. In Subtask-B, we finished fourth out of seven teams, and our submission was eleventh out of all the runs (16) submitted by the teams. In Subtask-C We finished fifth out of six teams, and our submission was fourteenth out of all the runs(15) submitted by the teams. In Task 3, we can observe that the first submission for subtask-A, which employed fast.ai, had a higher F_1 score than the other entries, which used multilingual models. The model was fine-tuned, resulting in modest variations ranging from 0.739 to 0.907. Our model underperformed in the other subtasks, which might be owing to overfitting the model caused by the unbalanced dataset, albeit we did apply data augmentation.

For each subtask, we have submitted all the various submissions. Following are descriptions of each run.

1. **submission-task1-1:** We have used mBERT for German-English code-mixed data. The maximum sequence length was 256 tokens, and the batch size of 32. For Hindi-English code-mixed data, we have used XLM-Roberta. The maximum sequence length was 512 tokens, and the batch size of 16. (Macro F1: 0.6270)
2. **submission-task1-2_t:** We have used GermanBERT for German-English code-mixed data. The maximum sequence length was 128 tokens, and the batch size of 32. For Hindi-English code-mixed data, we have used XLM-Roberta. The maximum sequence length was 512 tokens, and the batch size of 16. (Macro F1: 0.6192)
3. **submission-task2-1_t:** We have used XLM-Roberta for this task. The maximum sequence length was 512 tokens, the batch size of 16, and the epoch was 10. (Macro F1: 0.439)
4. **submission-task2-2:** We have used mBERT for this task. The maximum sequence length was 512 tokens, and the batch size of 16. We have frozen all the parameters of the pre-trained model and then used early stopping criteria. To avoid class imbalance we have used ”classweight=balanced.” (Macro F1: 0.307)
5. **submission-task2-3:** We have used mBERT for this task. The maximum sequence length was 512 tokens, and the batch size of 16. We have a Focal loss function here. (Macro F1: 0.387)
6. **submission-task2-4:** We have used XLM-Roberta for this task. The maximum sequence length was 512 tokens, and the batch size of 16. (Macro F1: 0.392)
7. **submission-task2-5:** We have used XLM-Roberta for this task. The maximum sequence length was 512 tokens, the batch size of 32, and the epoch was 2. (Macro F1: 0.6093)

Table 3
Evaluation results on test data and rank list (Submission number in bracket)

Language	Task	Team Name	F_1	P	R	Rank in Top	Rank in All
ICHCL	Task 1	nlplab_isi	.708	.712	.709	1 / 13	1 / 42
		irlab@iitbhu (1)	.619	.623	.620	-	16 / 42
		irlab@iitbhu (2)	**.627**	**.630**	**.628**	6 / 13	14 / 42
	Task 2	ub-cs	.493	.521	491	1 / 11	1 / 26
		irlab@iitbhu (1)	**.439**	**.553**	**.443**	7 / 11	10 / 26
		irlab@iitbhu (2)	.307	.371	.397	-	21 / 26
		irlab@iitbhu (3)	.387	.394	.396	-	17 / 26
		irlab@iitbhu (4)	.392	.419	.406	-	16 / 26
		irlab@iitbhu (5)	.375	.392	.395	-	18 / 26
Marathi	Task 3A	ssncse-nlp	.974	.975	.974	1 / 10	1 / 28
		irlab@iitbhu (1)	**.935**	**.935**	**.935**	6 / 10	9 / 28
		irlab@iitbhu (2)	.543	.543	.543	-	26 / 28
		irlab@iitbhu (3)	.907	.909	.907	-	13 / 28
		irlab@iitbhu (4)	.838	.846	.840	-	19 / 28
		irlab@iitbhu (5)	.739	.776	.748	-	24 / 28
	Task 3B	hate-busters	.920	.911	.934	1 / 7	1 / 16
		irlab@iitbhu (1)	**.535**	**.483**	**.610**	4 / 7	11 / 16
	Task 3C	satlab	.960	.936	.989	1 / 6	1 / 15
		irlab@iitbhu (1)	**.289**	**.285**	**.421**	5 / 6	14 / 15

8. **submission-task3a-1:** We have used XLM-Roberta for this task. The maximum sequence length was 128 tokens, the batch size of 16, and the epoch was 4. (Macro F1: 0.935)
9. **submission-task3b-1:** We have used XLM-Roberta for this task. The maximum sequence length was 128 tokens, the batch size of 16, and the epoch was 8. (Macro F1: 0.535)
10. **submission-task3c-1:** We have used XLM-Roberta for this task. The maximum sequence length was 128 tokens, the batch size of 16, and the epoch was 5. (Macro F1: 0.289)

6. Conclusion

In this paper, we have presented the system submitted by the IRLab@IITBHU team to the HASOC 2022 - Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages shared task at FIRE 2022. Our system is based on fine-tuning state of the art transformer models like XLM Roberta and German BERT to categorize tweets in Marathi and Hinglish Codemix and German language. Pre-trained bi-directional encoder representations using transformers outperform the traditional machine learning models.

7. Acknowledgements

We are thankful to the organizers of HASOC 2022 for providing the opportunity to work on this interesting and important task.

References

[1] R. Bose, R. K. Dey, S. Roy, D. Sarddar, Analyzing political sentiment using twitter data, in: Information and communication technology for intelligent systems, Springer, 2019, pp. 427–436.

[2] R. Ali, U. Farooq, U. Arshad, W. Shahzad, M. O. Beg, Hate speech detection on twitter using transfer learning, Computer Speech & Language 74 (2022) 101365.

[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] M. Zampieri, T. Ranasinghe, M. Chaudhari, S. Gaikwad, P. Krishna, M. Nene, S. Paygude, Predicting the type and target of offensive social media posts in marathi, Social Network Analysis and Mining 12 (2022) 77. URL: https://doi.org/10.1007/s13278-022-00906-8. doi:10.1007/s13278-022-00906-8.

[6] T. Ranasinghe, K. North, D. Premasiri, M. Zampieri, Overview of the HASOC subtrack at FIRE 2022: Offensive Language Identification in Marathi, in: Working Notes of FIRE 2022 - Forum for Information Retrieval Evaluation, CEUR, 2022.

[7] A. Saroj, S. Chanda, S. Pal, IRlab@IITV at SemEval-2020 task 12: Multilingual offensive language identification in social media using SVM, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 2012–2016. URL: https://aclanthology.org/2020.semeval-1.265.

[8] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & internet 7 (2015) 223–242.

[9] R. Kshirsagar, T. Cukuvac, K. McKeown, S. McGregor, Predictive embeddings for hate speech detection on twitter, arXiv preprint arXiv:1809.10644 (2018).

[10] B. Gambäck, U. K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the first workshop on abusive language online, 2017, pp. 85–90.

[11] Z. Zhang, D. Robinson, J. Tepper, Detecting hate speech on twitter using a convolution-gru based deep neural network, in: European semantic web conference, Springer, 2018, pp. 745–760.

[12] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Effective hate-speech detection in twitter data using recurrent neural networks, Applied Intelligence 48 (2018) 4730–4742.
[13] A. S. Saksesi, M. Nasrun, C. Setianingsih, Analysis text of hate speech detection using recurrent neural network, in: 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), IEEE, 2018, pp. 242–248.
[14] S. S. Syam, B. Irawan, C. Setianingsih, Hate speech detection on twitter using long short-term memory (lstm) method, in: 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), IEEE, 2019, pp. 305–310.
[15] A. R. Isnain, A. Sihabuddin, Y. Suyanto, Bidirectional long short term memory method and word2vec extraction approach for hate speech detection, IJCCS (Indonesian Journal of Computing and Cybernetics Systems) 14 (2020) 169–178.
[16] S. Chanda, S. Ujjwal, S. Das, S. Pal, Fine-tuning pre-trained transformer based model for hate speech and offensive content identification in english, indo-aryan and code-mixed (english-hindi) languages, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2021.
[17] M. Mozafari, R. Farahbakhsh, N. Crespi, A bert-based transfer learning approach for hate speech detection in online social media, in: International Conference on Complex Networks and Their Applications, Springer, 2019, pp. 928–940.
[18] F. M. Plaza-del Arco, M. D. Molina-González, L. A. Urena-López, M. T. Martín-Valdivia, Comparing pre-trained language models for spanish hate speech detection, Expert Systems with Applications 166 (2021) 114120.
[19] P. Chiril, F. Benamara, V. Moriceau, M. Coulomb-Gully, A. Kumar, Multilingual and multitarget hate speech detection in tweets, in: Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019), ATALA, 2019, pp. 351–360.
[20] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An italian twitter corpus of hate speech against immigrants, in: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 2018.
[21] S. Jaki, T. De Smedt, Right-wing german hate speech on twitter: Analysis and automatic detection, arXiv preprint arXiv:1910.07518 (2019).
[22] C. Baziotis, N. Pelekis, C. Doulkeridis, Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 747–754.
[23] 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, arXiv preprint arXiv:1911.02116 (2019).
[24] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, arXiv preprint arXiv:1808.06226 (2018).