RNN's VS TRANSFORMERS : Training language models on deficit datasets

Abhishek Kumar Gautam[1], B Bharathi[2]

[1] *Department of Computer science, Indian Institute of Information Technology Una, Himachal Pradesh, India*
[2] *Department of CSE, Sri Siva Subramaniya Nadar College of Engineering,Tamil Nadu, India*

Abstract

The concept of content moderation is as old as the online social media itself, the goal is to prevent any hate speech, comments etc to happen on the platform so as to keep the online social environment friendly and sane. With an exponentially increasing number of people on social media content moderation is a difficult task as such in the modern era we make use of specialised tools such as AI and NLP. In non-native English spoken countries, social media texts are mostly in code mixed form. This paper discusses the work put by SSNCSE_NLP in HASOC offensive language identification on multilingual codemixed text tasks of FIRE 2021. In this paper we have put a detailed comparison on the performance of several RNN's based models with transformers based BERT architecture by varying the essential hyperparameters when training on a smaller dataset for tasks like sentiment analysis. We achieved an F1 score of 72.47% in task 1 and 69.2% ,61.5% in task2 Tamil and Malayalam respectively on the test set from our best evaluated model.

Keywords

Offensive content, Dravidian languages, RNN, LSTM

1. Introduction

Every small or big brand wants to put their product into as many hands as possible and easily accessible in their own native languages this, in combination with the reach of internet has resulted into massive expanse of rich diverse user groups, online content moderation in those native languages along with the mixed languages that the user group speaks is hence necessary. As codemixed languages consist of bilingual, trilingual or more languages in combination with symbols and emojis it's difficult to train efficient models. With recent developments in sequence processing models and Transformer[1] based architectures it is far easier to train models in these mixed languages sets. The review of code mixed research and challenges involved in speech and language processing is discussed in [2]. Ensemble approach for offensive identification were discussed [3]. Multilingual BERT based transformer models are used for offensive language identification task [4]. In this paper we have compared training LSTM[5] based architectures with transformers based BERT[6] when training on smaller datasets on codemixed Dravidian languages Malayalam and Tamil mixed with English. Machine learning based approaches for

FIRE 2021: Forum for Information Retrival Evaluation, December 13-17, 2021, India
19105@iiitu.ac.in (A.K. Gautam); bharathib@ssn.edu.in (B. Bharathi)
https://www.ssn.edu.in/staff-members/dr-b-bharathi/ (B. Bharathi)
0000-0002-8916-6351 (A.K. Gautam); 0000-0001-7279-5357 (B. Bharathi)

Task - 1	“ தம்பி... நீங்க எல்லாம் YouTube-கு புதுசா?? எந்த video வ இருந்தாலும் மொதல்ல likes தான் update ஆகும் அதுக்கப்பறம் தான் views update ஆகும்...puthusa jio sim vangi puthusa YouTube la pakravanga ellam konjam silent ah iruntha nalla irukum”
Task -2 – (Tamil – English)	Take it this thevidiya Kandipa indha page admin Oru Mutual Punda Vijay fan ha Dhan erupan
Task – 2 (Malayalam – English)	“ USER cheruparamadathil than thinnunnath alla pinarayi thinnunnath pinarayikk oru barber venam mudi kalayan jacob thomas vannal aa joli ayale elpikkum ”.

Figure 1: Example sentences of task 1 and task 2.

offensive language identification are described in [7]. Tamil and Malayalam belong to the Dravidian language family spoken mainly in south India, Sri Lanka, and Singapore.

The paper is organized as follows: The dataset descriptions are given in Section2.1 Section 2.3 details the experimental setup and various features used for this task. Section 3 provides a subjective analysis and comparison of the performance of various models on the development and test data. Finally, Section 4 concludes the paper.

2. Proposed work

2.1. Dataset analysis and task description

The primary goal of this shared task is to detect offensive language of the code-mixed dataset of comments/posts in Dravidian Languages (Malayalam-English and Tamil-English) collected from social media [8][9]. The comment/post may contain more than one sentence but the average sentence length of the corpora is 1. Each comment/post is annotated with offensive language label at the comment/post level[10]. This dataset also has class imbalance problems depicting real-world scenarios. The HASOC Dravidan dataset had 2 tasks, for the first task, we were given with a message-level label classification task. Given a YouTube comment in Tamil, the model had to classify it into offensive or not-offensive. For the second task, Given a tweet in codemixed Tamil and Malayalam, systems have to classify it into offensive or not-offensive. Example sentences of task 1 and task 2 is given in Fig.1.

2.2. Preprocessing

The datasets consisted of Dravidian languages Tamil and Malayalam codemixed with English words, symbols and emojis. The dataset was parsed to generate word level tokens then generate characters to separate out non-UTF-8 charset also emojis were removed from the dataset obtained from the charset, later the word level tokens were directly parsed into LSTM based networks while the clean text were parsed separately to generate BERT-tokens for training the trans-

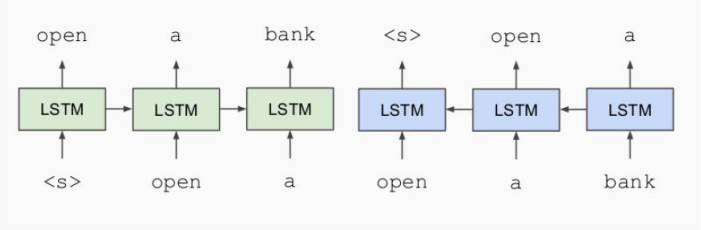

Figure 2: use of separate left-to-right (green LSTM blocks) and right-to-left (blue LSTM blocks) models in ELMo [12]

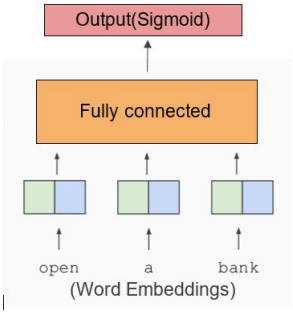

Figure 3: ELMo for sequence classification [12]

former architecture this was done so as to create proper tokens for Tamil in English(Tanglish) and Malayalam in English(Manglish) datasets provided in task-2 while pre-trained embeddings from Indic BERT[11] was used for task-1.

2.3. Experiments

For natural language modelling several LSTM architectures including ELMo[7] were tested along with state-of-art transformer model BERT. Considering lack of vocabulary or unknown tokens in embeddings for Tanglish and Manglish datasets of task-2 the models were trained from scratch for task-2. The LSTM based RNN architectures were created and trained in Tensorflow while the transformer architecture BERT was trained in Pytorch using Huggingface transformers library. The Jupyter notebooks for both training tasks are available here.

2.3.1. ELMo model

ELMo is a LSTM based architecture that leverages bi-directionality[11] of natural language models by using two separate LSTM layers running left-to-right and right-to-left in bidirectional wrapper and shallow concatenating the outputs. Since there weren't any multilingual models on ELMo for Indian languages we ended up training it from scratch and achieved an accuracy of 82.7% on validation set.

ELMo utilizes LSTM's it shares same hyper-parameters as them namely :

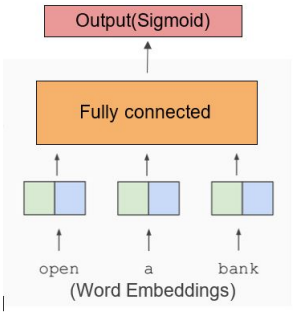

Figure 4: Fine-tuning BERT encoder transformer [6]

1. Number of units : number of LSTM units or output dimension.
2. Dropout : dropout rate(0-1) for outputs.
3. Recurrent dropout : dropout rate(0-1) for recurrent output . ELMo architecture in contrast to the fairly small dataset could be trained from scratch for better accuracy.

2.3.2. BERT model

BERT is a transformer[4] based architecture which has proven to be fit for a wide variety of tasks[5] it uses self attention to generate understanding within the network. Task-1 had plain Tamil text code mixed with English words so fine-tuning multilingual model Indic BERT gave a score of 81% on validation set. For task-2 we trained the models from scratch , since the maximum sentence length was found to be 91, embedding size of 128 was used with mini-BERT configuration to pre-train the model on text. Pre-training was done on the entire set on Masked language modelling then the model was fine-tuned for classification on downstream tasks.

BERT models are implemented in pytorch and utilises hugging face transformers API to create and train models, while the ELMo architecture was implemented using tensorflow and was trained from scratch all the notebooks associated with training of models are available in the link [1]

3. Performance analysis

The performance of the proposed approach using BERT model is given in Table1. From 1, it has been noted that fine-tuning multilingual model Indic BERT gave a score of 81% on validation set.

The performance of the proposed system using ELMO model is given in Table 2.

In Table 2 BiLSTMu refers to BiLSTM with side-by-side stacked uni-directional(left-to-right) LSTM. BiLSTMd refers to BiLSTM with separate left-to-right and right-to-left LSTM stacked, outputs shallow concatenated and fed to fully connected layers for classification.

[1]https://github.com/Abhishek-krg/Multilingual-codemixed-language-classification

Table 1
Performance of proposed system using BERT model with validation data

Attention heads	Hidden layers	Hidden size	Embeddings	Accuracy (in %)
2	12	128	128	80.66
2	12	256	128	81.2
4	12	128	128	80.38
2	12	256	64	70.51
4	12	256	128	81.06
2	12	512	128	79.37

Table 2
Performance of proposed system using LSTM models with validation data

Architecture	Units	Dropout (in %)	Recurrent-dropout (in %)	Accuracy (in %)
Unidirectional-LSTM	64	20	20	80
BiLSTMu	32	20	20	81
BiLSTMu	64	20	20	81.6
BiLSTMd	32	20	20	82.5
BiLSTMd	64	20	20	82.5
BiLSTMd	32	25	10	82.7

Table 3
Performance of proposed system using XLM-RoBERTa models with validation data

Model	Accuracy in (%)
XLM-RoBERTa-base	80.5
XLM-RoBERTa-large	81.2

Table 4
Performance of proposed system using test data

Task	Precision	Recall	F1 score
Task-1 Tamil	0.747	0.725	0.735
Task-2 Tamil-English	0.615	0.607	0.61
Task-2 Malayalam -English	0.692	0.678	0.683

From Table 2, it has been noted that BiLSTMd model with 32 units achieves highest accuracy of 82. 7%. Considering the performances of multilingual LM, we have experimented XLM-Roberta. The results are tabulated in Table 3.

The performance of the proposed system using the test data is given in Table 4.

4. Conclusion

In this paper, we proposed offensive language identification using Dravidian code-mixed text using ELMO and BERT models. From the performance metrics above it is clear that BERT despite being a far better architecture couldn't achieve expected results while the BiLSTMd architecture gave better results on HASOC dataset. This could be due to the reason that BERT is a very dense model and requires huge-datasets to train on while LSTM based RNN architectures on the other hand can achieve better results on simpler classification tasks.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[2] S. Sitaram, K. R. Chandu, S. K. Rallabandi, A. W. Black, A survey of code-switched speech and language processing, 2020. arXiv:1904.00784.

[3] D. Saha, N. Paharia, D. Chakraborty, P. Saha, A. Mukherjee, Hate-alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 270–276. URL: https://www.aclweb.org/anthology/2021.dravidianlangtech-1.38.

[4] S. M. Jayanthi, A. Gupta, Sj_aj@dravidianlangtech-eacl2021: Task-adaptive pre-training of multilingual bert models for offensive language identification, 2021. arXiv:2102.01051.

[5] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, LSTM: A search space odyssey, CoRR abs/1503.04069 (2015). URL: http://arxiv.org/abs/1503.04069. arXiv:1503.04069.

[6] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[7] A. B. Nitin Nikamanth, B. Bharathi, Ssncse_nlp@hasoc-dravidian-codemix-fire2020: Offensive language identification on multilingual code mixing text, in: Working Notes of FIRE 2020- Forum for Information Retrieval Evaluation, CEUR, 2020, pp. 370–376.

[8] B. R. Chakravarthi, R. Priyadharshini, N. Jose, A. Kumar M, T. Mandl, P. K. Kumaresan, R. Ponnusamy, H. R L, J. P. McCrae, E. Sherly, Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, Kyiv, 2021, pp. 133–145. URL: https://aclanthology.org/2021.dravidianlangtech-1.17.

[9] R. Priyadharshini, B. R. Chakravarthi, S. Thavareesan, D. Chinnappa, T. Durairaj, E. Sherly, Overview of the dravidiancodemix 2021 shared task on sentiment detection in tamil, malayalam, and kannada, in: Forum for Information Retrieval Evaluation, FIRE 2021, Association for Computing Machinery, 2021.

[10] B. R. Chakravarthi, P. K. Kumaresan, R. Sakuntharaj, A. K. Madasamy, S. Thavareesan,

P. B, S. Chinnaudayar Navaneethakrishnan, J. P. McCrae, T. Mandl, Overview of the HASOC-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.

[11] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, P. Kumar, IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages, in: Findings of EMNLP, 2020.

[12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, CoRR abs/1802.05365 (2018). URL: http://arxiv.org/abs/1802.05365. arXiv:1802.05365.