Dharwad
ज्ञानेन विकासः

INDIAN INSTITUTE OF
INFORMATION
TECHNOLOGY

Indian Institute of Information Technology Dharwad
Department of Computer Science Engineering

Mini Project - II Report

Building a framework for Hostile Content Detection in Marathi Tweets

Submitted by :

Roll No	Name of Student
19BCS047	K.V.S Bharadwaj
19BCS083	P. Satwik
19BCS084	P. Rutwik
19BCS093	R. Dhatri Kiran Reddy

Under the guidance of

Dr. Sunil Saumya

1. Introduction

Hostile Content Detection in Social Media posts is a herculean task that has become highly relevant and pivotal in the current scenario. Therefore, it is essential to build a robust framework which can detect such posts that might prove to be significant in reducing the hate induced by such posts.

The Social media platform we wish to focus on this project is Twitter. In accordance with the subject of Hate Speech, Offensive or Profane Content, Twitter amongst its contemporaries is highly relevant for such a study due to the new reforms it has brought into its policies to encourage free speech, as quoted by its new chief.

As far as the language of interest we chose is Marathi, it is an Indo-Aryan language predominantly spoken in the Indian state of Maharashtra by around 83 million people, making it the third-largest spoken language in India.

In this we explore a multilingual pre trained BERT transformer model for hostile content detection hostile content detection in Marathi tweets. It is a binary hate speech detection dataset in Marathi. We provide a detailed analysis of the performance of multilingual BERT model trained along with Bidirectional LSTM model for classification.

Our contributions to the field of Research is that we explore a multi model architecture to get better results in classification, where we created three sub models to classify all of the 3 subtasks. To the best of our knowledge, this is the first work to explore such multi-model architecture on the twitter Bert models.

2. Related Work

Hate speech detection is seen as a major issue, and several measures have been undertaken to regulate it. English text analysis requires a great amount of labor. However, new efforts have been made to broaden studies into regional languages such as Marathi.

In [11], the Marathi Offensive Language Dataset (MOLD) was presented, which included over 2,500 annotated tweets categorized as offensive or non offensive. It is regarded as the first dataset for identifying foul language in Marathi. They also assessed the performance of many classic machine learning models as well as deep learning models (such as LSTM) trained on MOLD.

The MIMCT to detect offensive (Hate or Abusive) Hinglish tweets using the planned Hinglish Offensive Tweet dataset was given in [24]. The multi-channel CNN-LSTM model was used for sentiment analysis.

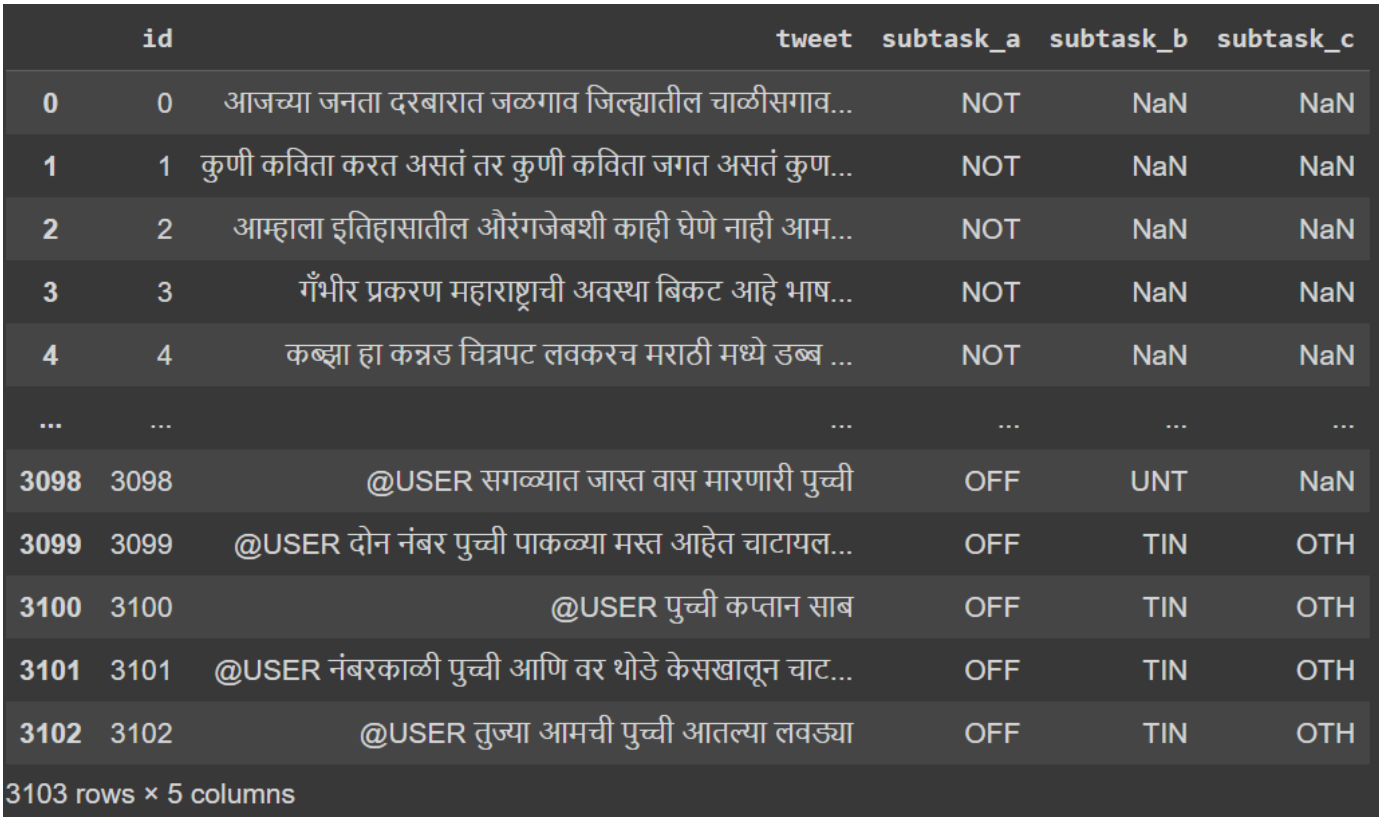
The authors of [20] provided a dataset of over 16000 Marathi tweets that were manually classified as good, negative, or neutral. They also developed a guideline for categorizing phrases based on their emotional content. CNN, BiLSTM, and BERT models were used in the analysis.

[4] created a Hindi-English code-mixed corpus from tweets posted online over a five-year period. Twitter python API was used to scrape tweets by choosing certain hashtags and phrases from political events, public protests, riots, and so on.

3. Methodology

3.1. Dataset

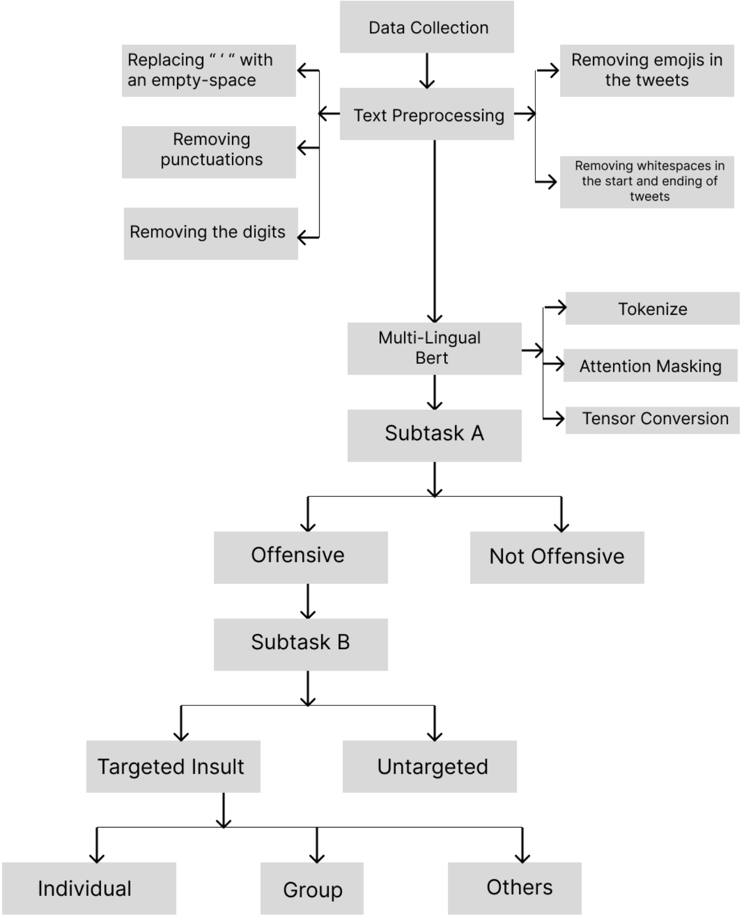
We fine tune our model on the HASOC 2022 dataset. This dataset consists of tweets from social media in Marathi. The dataset is divided into three subtasks: Subtask A, Subtask B, and Subtask C. The data points in subtask A are labeled as offensive and not offensive; those in subtask B are labeled as targeted insult and untargeted; the remaining in subtask C are labeled as individual, group, and others. We train the model with 70% of the data and validate it with the remaining 30%. The data comprises a total of 3103 data points, of which 2041 are offensive and 1062 are non-offensive.



	id	tweet	subtask_a	subtask_b	subtask_c
0	0	आजच्या जनता दरबारात जळगाव जिल्ह्यातील चाळीसगाव...	NOT	NaN	NaN
1	1	कुणी कविता करत असतं तर कुणी कविता जगत असतं कुण...	NOT	NaN	NaN
2	2	आम्हाला इतिहासातील औरंगजेबशी काही घेणे नाही आम...	NOT	NaN	NaN
3	3	गँभीर प्रकरण महाराष्ट्राची अवस्था बिकट आहे भाष...	NOT	NaN	NaN
4	4	कब्झा हा कन्नड चित्रपट लवकरच मराठी मध्ये डब्ब ...	NOT	NaN	NaN
...	...	...	...	...	...
3098	3098	@USER सगळ्यात जास्त वास मारणारी पुच्ची	OFF	UNT	NaN
3099	3099	@USER दोन नंबर पुच्ची पाकळ्या मस्त आहेत चाटायल...	OFF	TIN	OTH
3100	3100	@USER पुच्ची कप्तान साब	OFF	TIN	OTH
3101	3101	@USER नंबरकाळी पुच्ची आणि वर थोडे केसखालून चाट...	OFF	TIN	OTH
3102	3102	@USER तुज्या आमची पुच्ची आतल्या लवड्या	OFF	TIN	OTH

3103 rows × 5 columns

3.2. Flow Diagram



3.3. Model Architecture

3.3.1. Data preprocessing and tokenizing

We chose to use the BERT-based model as they have shown good results for text classification. Pre-training these models on huge datasets have proven to yield better results on downstream classification tasks in the same language. In our scenario, we have to use the mBERT, multilingual Bert model, which is a model released along with BERT, supporting 104 languages. It is basically just BERT trained on text across various languages.

First, we preprocessed the data to gain much better results on the classification task. We performed cleaning operations to ensure the ideal conditions of the data. The provided dataset had emojis, various punctuations, empty spaces, digits etc. which all make it difficult for the model to classify. Our preprocessing methods cleaned all such unnecessary data to make it easier for tokenizing.

Then, all the tweets were tokenized by the mBERT tokenizer before being used by the model. The tokenized text will be used as the input for the model backbone.

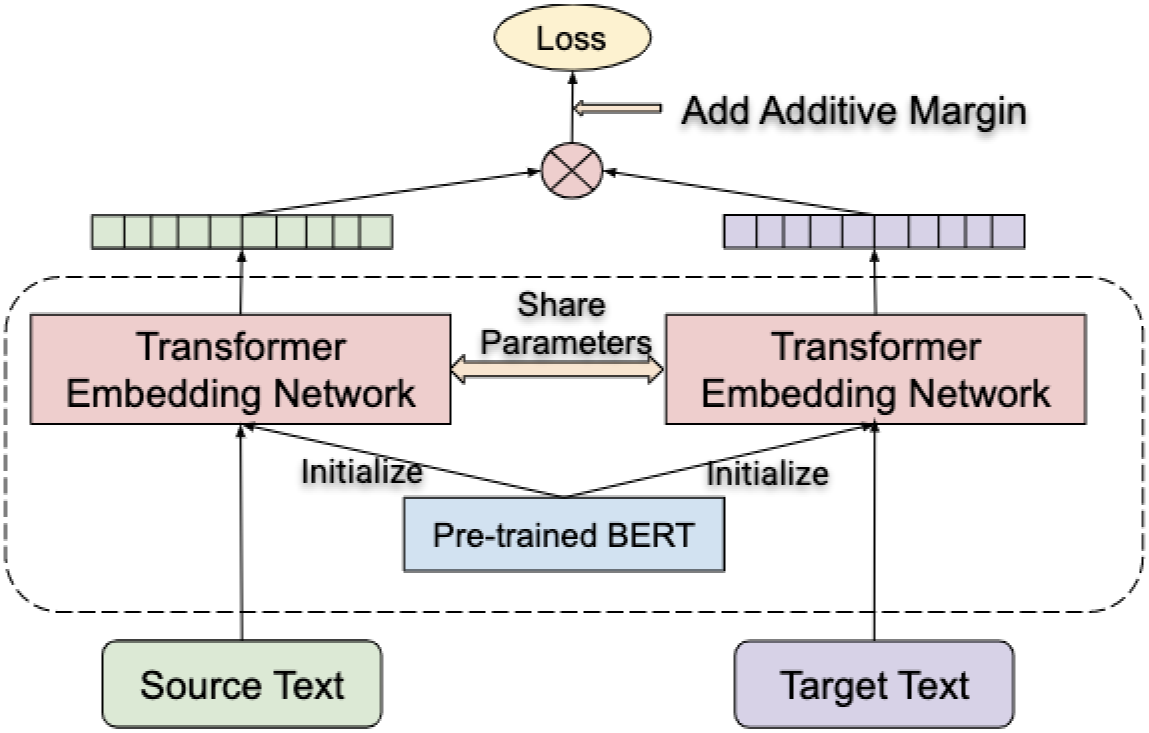


Fig.: Language-agnostic BERT embedding (mBERT)

3.3.2. Model Architecture

We have chosen to use the bidirectional LSTM, commonly known as biLSTM model, which is a sequence processing model with a special architecture.

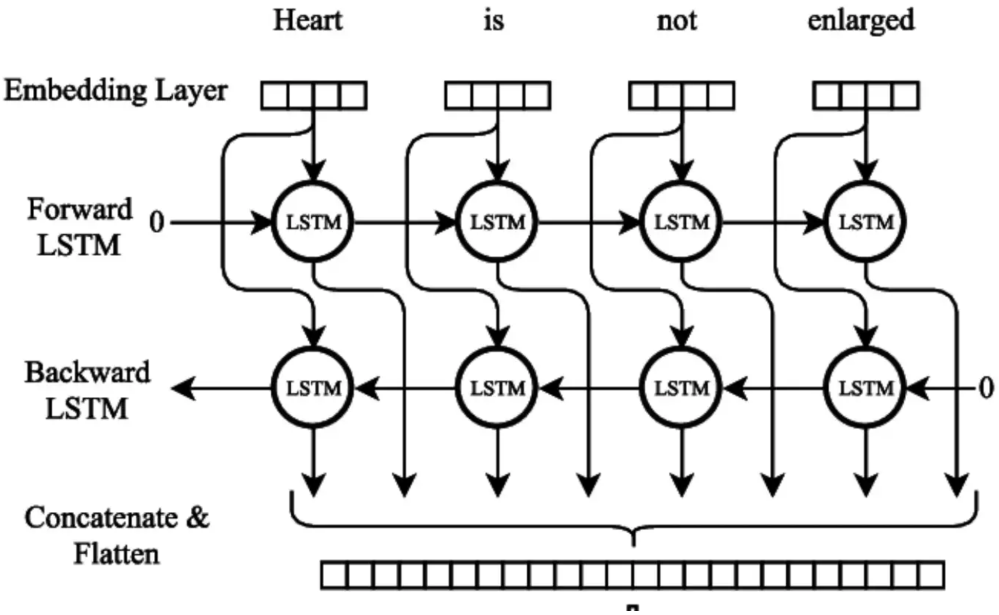
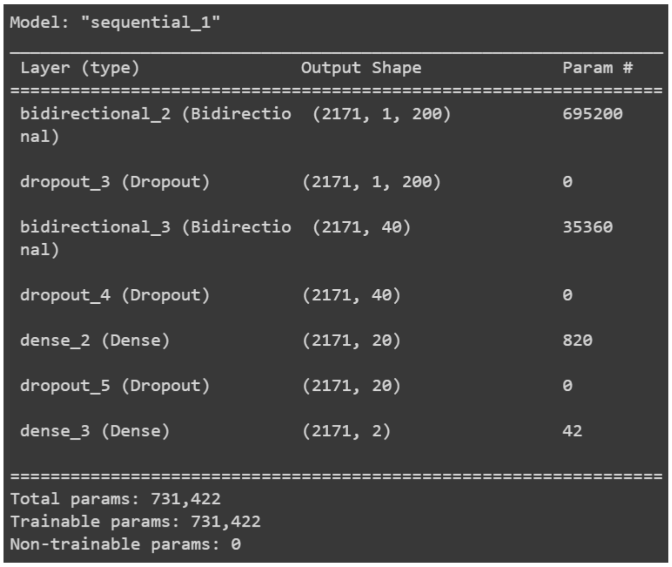


Fig.: biLSTM Architecture

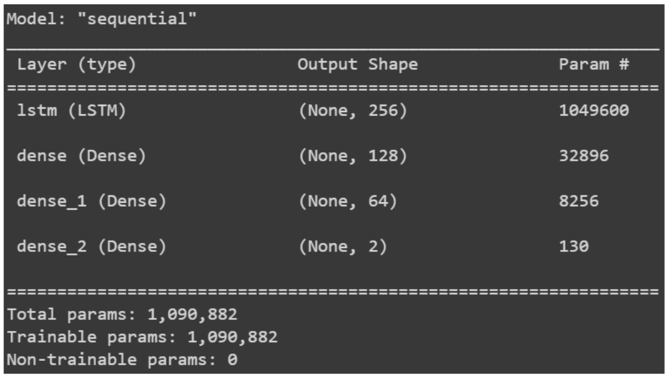
As you can see above, there are two LSTM models: one working forward and the other in the reverse direction. Due to this, biLSTMs enable additional training by passing the text sequence twice. Therefore, the biLSTM model completes more training on a given dataset than LSTM, which helps it offer better predictions.

As mentioned above, we have used a multi-model architecture. All three subtasks have their own different architectures. For subtask A, we have used the biLSTM model; whereas for subtask B and C, we have used LSTM.



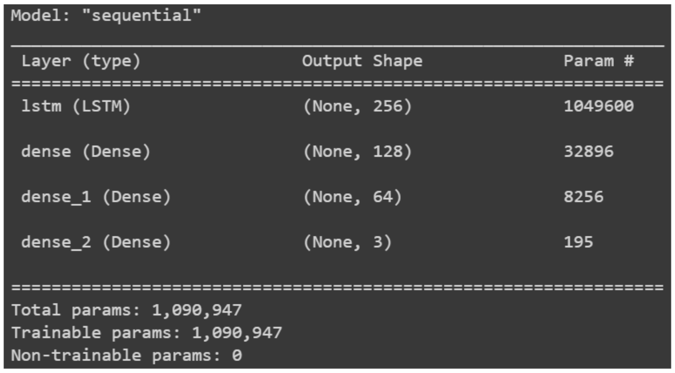
```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bidirectional_2 (Bidirectio  (2171, 1, 200)           695200
 nal)

 dropout_3 (Dropout)         (2171, 1, 200)            0

 bidirectional_3 (Bidirectio  (2171, 40)               35360
 nal)

 dropout_4 (Dropout)         (2171, 40)                0

 dense_2 (Dense)             (2171, 20)                820

 dropout_5 (Dropout)         (2171, 20)                0

 dense_3 (Dense)             (2171, 2)                 42

=================================================================
Total params: 731,422
Trainable params: 731,422
Non-trainable params: 0
```

Fig.: Model Architecture for Subtask A



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 256)               1049600

 dense (Dense)               (None, 128)               32896

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 2)                 130

=================================================================
Total params: 1,090,882
Trainable params: 1,090,882
Non-trainable params: 0
```

Fig.: Model Architecture for Subtask B



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 256)               1049600

 dense (Dense)               (None, 128)               32896

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 3)                 195

=================================================================
Total params: 1,090,947
Trainable params: 1,090,947
Non-trainable params: 0
```

Fig.: Model Architecture for Subtask C

Results

We hereby present our results on the HASOC 2022 dataset training split. We report macro-F1 scores to get a clear idea of the performance of models. Results of all the models are presented below:

Subtask A

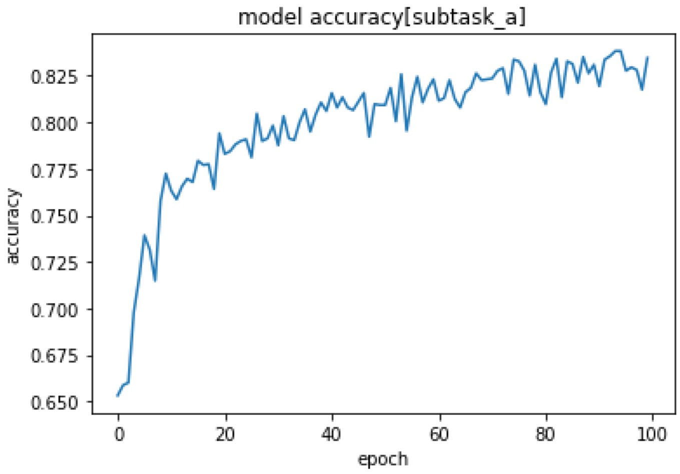
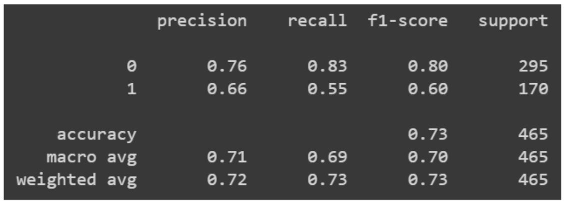


Fig.: Plot for model accuracy



```
              precision    recall  f1-score   support

           0       0.76      0.83      0.80       295
           1       0.66      0.55      0.60       170

    accuracy                           0.73       465
   macro avg       0.71      0.69      0.70       465
weighted avg       0.72      0.73      0.73       465
```

Fig.: Performance matrix for subtask A

Subtask B

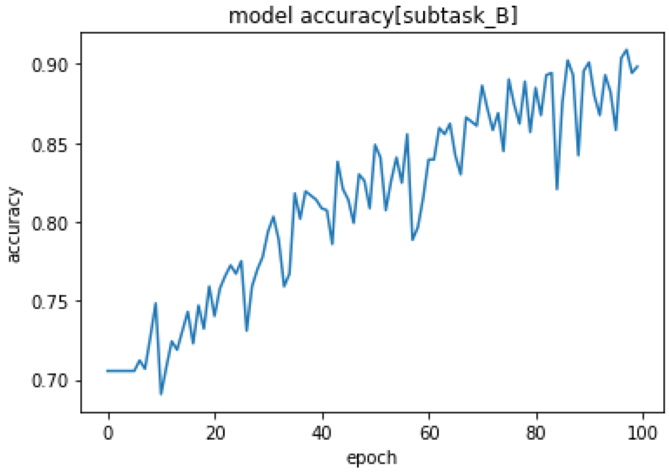
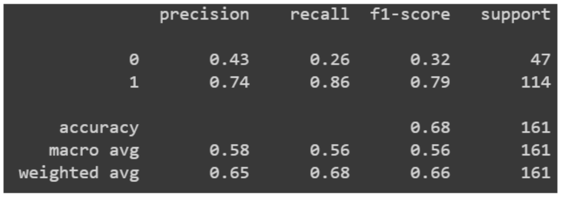


Fig.: Plot for model accuracy



```
              precision    recall  f1-score   support

           0       0.43      0.26      0.32        47
           1       0.74      0.86      0.79       114

    accuracy                           0.68       161
   macro avg       0.58      0.56      0.56       161
weighted avg       0.65      0.68      0.66       161
```

Fig.: Performance matrix for subtask B

Subtask C

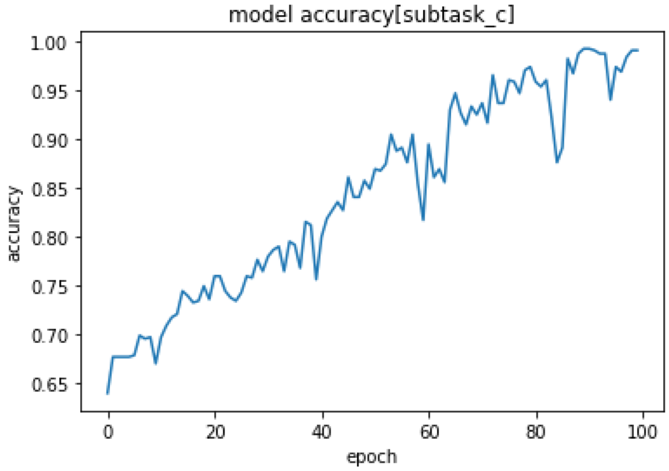
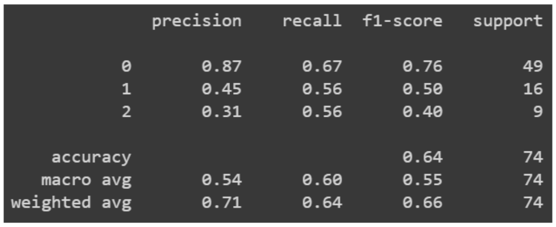


Fig.: Plot for model accuracy



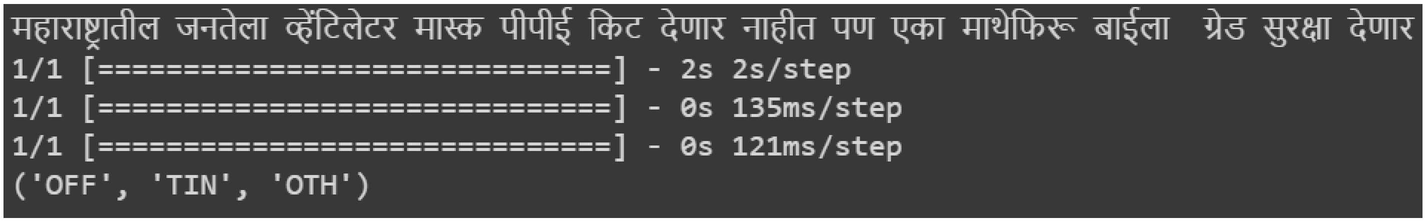
```
              precision    recall  f1-score   support

           0       0.87      0.67      0.76        49
           1       0.45      0.56      0.50        16
           2       0.31      0.56      0.40         9

    accuracy                           0.64        74
   macro avg       0.54      0.60      0.55        74
weighted avg       0.71      0.64      0.66        74
```

Fig.: Performance matrix for subtask B

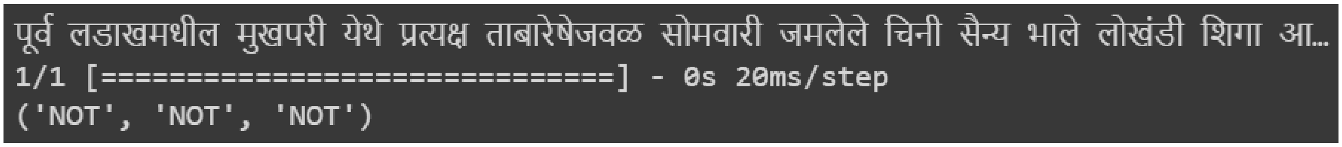
Analysis

One of the key observations from the above results is that the multi-model architecture gives way better results than using a single model for all the three subtasks. We have tested this by passing a single sentence from the test dataset to model. The dataset does not contain any labels, and to get a single output instead of three outputs from three different models; we have written a function that combines the weights of all the three models and gives us the output from all the three models at once.

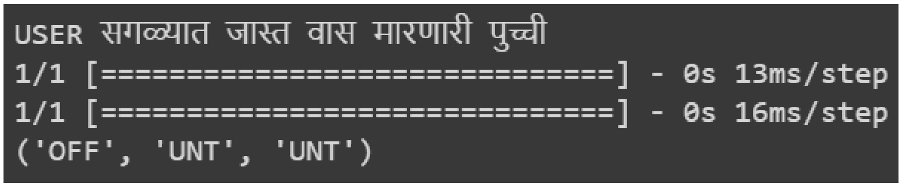
Testing on single sentences from the test dataset :



```
महाराष्ट्रातील जनतेला व्हेंटिलेटर मास्क पीपीई किट देणार नाहीत पण एका माथेफिरू बाईला  ग्रेड सुरक्षा देणार
1/1 [==============================] - 2s 2s/step
1/1 [==============================] - 0s 135ms/step
1/1 [==============================] - 0s 121ms/step
('OFF', 'TIN', 'OTH')
```

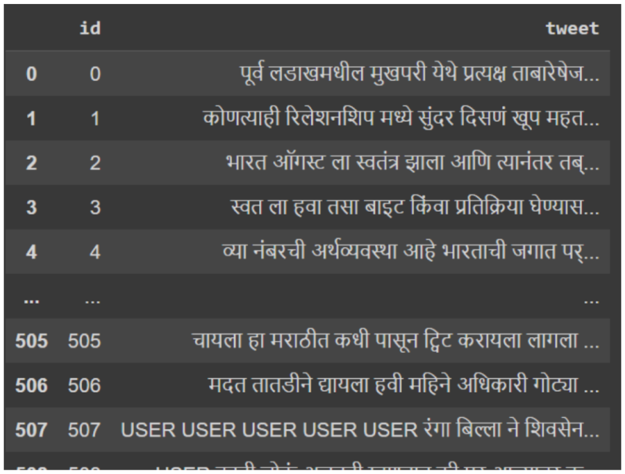


```
पूर्व लडाखमधील मुखपरी येथे प्रत्यक्ष ताबारेषेजवळ सोमवारी जमलेले चिनी सैन्य भाले लोखंडी शिगा आ…
1/1 [==============================] - 0s 20ms/step
('NOT', 'NOT', 'NOT')
```



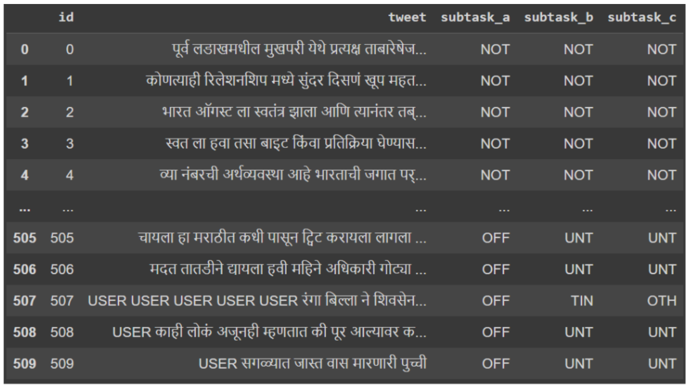
```
USER सगळ्यात जास्त वास मारणारी पुच्ची
1/1 [==============================] - 0s 13ms/step
1/1 [==============================] - 0s 16ms/step
('OFF', 'UNT', 'UNT')
```

After performing analysis on the test dataset, we have received the outputs for the whole dataset as presented below:



	id	tweet
0	0	पूर्व लडाखमधील मुखपरी येथे प्रत्यक्ष ताबारेषेज...
1	1	कोणत्याही रिलेशनशिप मध्ये सुंदर दिसणं खूप महत...
2	2	भारत ऑगस्ट ला स्वतंत्र झाला आणि त्यानंतर तब्...
3	3	स्वत ला हवा तसा बाइट किंवा प्रतिक्रिया घेण्यास...
4	4	व्या नंबरची अर्थव्यवस्था आहे भारताची जगात पर्...
...	...	...
505	505	चायला हा मराठीत कधी पासून ट्विट करायला लागला ...
506	506	मदत तातडीने द्यायला हवी महिने अधिकारी गोट्या ...
507	507	USER USER USER USER USER रंगा बिल्ला ने शिवसेन...

Fig.: Test dataset before analysis



	id	tweet	subtask_a	subtask_b	subtask_c
0	0	पूर्व लडाखमधील मुखपरी येथे प्रत्यक्ष ताबारेषेज...	NOT	NOT	NOT
1	1	कोणत्याही रिलेशनशिप मध्ये सुंदर दिसणं खूप महत...	NOT	NOT	NOT
2	2	भारत ऑगस्ट ला स्वतंत्र झाला आणि त्यानंतर तब्...	NOT	NOT	NOT
3	3	स्वत ला हवा तसा बाइट किंवा प्रतिक्रिया घेण्यास...	NOT	NOT	NOT
4	4	व्या नंबरची अर्थव्यवस्था आहे भारताची जगात पर्...	NOT	NOT	NOT
...	...	...	...	...	...
505	505	चायला हा मराठीत कधी पासून ट्विट करायला लागला ...	OFF	UNT	UNT
506	506	मदत तातडीने द्यायला हवी महिने अधिकारी गोट्या ...	OFF	UNT	UNT
507	507	USER USER USER USER USER रंगा बिल्ला ने शिवसेन...	OFF	TIN	OTH
508	508	USER काही लोकं अजूनही म्हणतात की पूर आल्यावर क...	OFF	UNT	UNT
509	509	USER सगळ्यात जास्त वास मारणारी पुच्ची	OFF	UNT	UNT

Fig.: Test dataset after analysis

The dataset was also cross-validated manually by a marathi candidate to check the accuracy of the model. It is to be noted that the model predicted 95% of the tweets correctly when cross-validated.

Conclusion

In this report, our main target was to experiment with mBERT and the multi-model architecture for various subtasks in our dataset for the classification of hate speech in marathi tweets. We have tried to tackle this classification task with a variant of the LSTM classifier, biLSTM. A few of the most recent related works have shown that unconventional input representation such as word-embeddings could be useful for this task even with the conventional classifiers.

We also observed that hate speech detection is a difficult task to accomplish, and is data driven. We also saw that not all the datasets could help achieve good results, especially because of the nature of the datasets and annotation schemes. Hate speech also heavily depends on the culture and location of the text being encountered. Current study shows that a classical ML approach with fine-tuned feature engineering could compete with state-of-the-art deep neural network-based models and in many ways are better. Our findings also support this hypothesis.

In a nutshell, our efforts in this report can be seen as:

- We have tried to implement an ML model, Bi-LSTM for the task of hate speech detection in Marathi tweets.
- We have experimented with multiple text vectorization models for Indic languages like mBERT.

Acknowledgements

We would like to share our sincere gratitude to everyone who helped us achieve this project. During the implementation, we learned new ideas and mechanisms. We did face challenges while learning and executing but we are grateful to our mentors for helping us get over our difficulties and compile our ideas into working code.

We would like to thank our project guide, **Dr. Sunil Saumya** for mentoring us and giving us this opportunity to apprise ourselves with these new concepts. We would also thank Mr. Shankar Biradar for helping us with the implementation.

K.V.S Bharadwaj
P. Satwik
P. Rutwik
R. Dhatri Kiran Reddy

Sixth Semester (Jan-May 2022)
Indian Institute of Information Technology Dharwad

References

1. https://arxiv.org/pdf/2212.10039.pdf
2. https://arxiv.org/pdf/2203.13778.pdf
3. Gaikwad, S., Ranasinghe, T., Zampieri, M., Homan, C.M.: Cross-lingual offensive language identification for low resource languages: The case of marathi (2021)
4. Mathur, P., Sawhney, R., Ayyar, M., Shah, R.: Did you offend me? classification of offensive tweets in hinglish language. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 138–148 (2018)
5. Kulkarni, A., Mandhane, M., Likhitkar, M., Kshirsagar, G., Joshi, R.: L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 213–220 (2021)
6. Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., Shrivastava, M.: A dataset of hindi-english code-mixed social media text for hate speech detection. In: Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in Social media. pp. 36–41 (2018)