Automatic Detection of Religiously Abusive Text on Social Media using Deep Learning Techniques

Presented By
Alamgir Hossain
ID: 1604069
Dept. of CSE, CUET

Supervised By
Dr. Mohammed Moshiul Hoque
Professor
Dept. of CSE, CUET

Contents

Introduction

- Provocations on occasion, as well as the ease of access to social media, have led to huge dissatisfaction among online religious groups.

- Religious abuse is one of the most abundant categories among of abuse.

- Identifying and categorizing religious abuse manually is a difficult task.

- This kind of task is still in preliminary stage for low resource languages like Bengali.

Motivation

- No system is developed yet to identify and categorize religiously abusive Bengali texts.
- Such system is required to ensure security in the cyberspace
- The spread of toxicity over social media can be reduced.
- Develop resources and models for Bengali.

Challenges

- Lack of linguistic tools
- Scarcity of benchmark corpora
- Overlapping characteristics with correlated phenomena
- Dealing with a large number of different local words.

Task Description

- Hierarchical annotation schema to divide the corpus into two levels:
  - **(A)** Religious text identification
  - **(B)** Classification of religious texts

- **Level A: Religious Text Identification**

  - **Religious texts (RE):** religious belief, attack, incite or seek to harm an individual, group or community based on some criteria such as religious ideology.
  - **Non religious texts (NoRE):** do not contain any statement of religion or express hidden wish/intent to harm other religion.

Task Description(Cont.)

- **Level B: Classification of Religious Text**
  - **Religious Abusive (ReAb):** incite violence by attacking religion (Islam, Hindu, Catholic, etc.), religious organizations, or religious belief of a person or a community.
  - **Religious Non Abusive (ReNoAb):** normal religious ideology, belief, mannerly quote or peaceful statement that doesn't provoke religious belief of individuals or a community.

Text	Level A	Level B
আজকে বৃষ্টি হবেই হবে	Non-Religious	-
তোদের যে ধর্ম ! থু থু!!	Religious	Abusive
ভাই তর্ক না করি, একমাত্র আল্লাহই পারেন তাদেরকে হিদায়াত দিতে।	Religious	Non-Abusive

Previous Work

1. Abusive content detection in transliterated Bengali-English social media corpus[1] (Sazzed, 2021)

- Can detect abusive or not from any transliterated Bangla text.
- Used Support Vector Machine (SVM).

Limitations:

- Deals with YouTube comments only.
- Detect only two classes.

2. Multi-label hate speech and abusive language detection in Indonesian Twitter[2] (Ibrohim et al., 2021)

- Multi-label text classification for abusive language and hate speech
- Used Random Forest Decision Tree (RFDT).

Limitations:

- Worked only on Indonesian Tweets.

Previous Work(Cont.)

3. An abusive text detection system based on enhanced abusive and non-abusive word lists[4] (Lee et al., 2018)

- Enhanced abusive and non-abusive word lists.
- Used unsupervised learning of abusive words.

Limitations:

- Used a word list that can be used for multiple meanings in the real world.
- Detect only two classes.

4. Identify Abusive and Offensive Language in Indonesian Twitter using Deep Learning Approach[5] (Ibrohim et al., 2019)

- Implemented deep learning approach to identify abusive language.
- LSTM with FastText performed better.

Limitations:

- Worked only on Indonesia Tweets.
- Identify only two categories.

Contributions

- Developed a Religious Text Corpus containing Bengali texts. Hierarchical annotation schema uses to classify Religious texts into abusive and non-abusive classes.

- Prepared a model to identify religiously abusive texts by investigating several deep learning models.

- Evaluated model performances to find the appropriate model for Bengali Religious Texts.

Dataset Description

Level	Class	Train	Test
Identification	Religious	2180	727
	Non-religious	826	275
Classification	Religious Abusive	1573	530
	Religious Non-abusive	607	197

Table 1 : Level-wise number of sample texts in each category

Annotation Schema

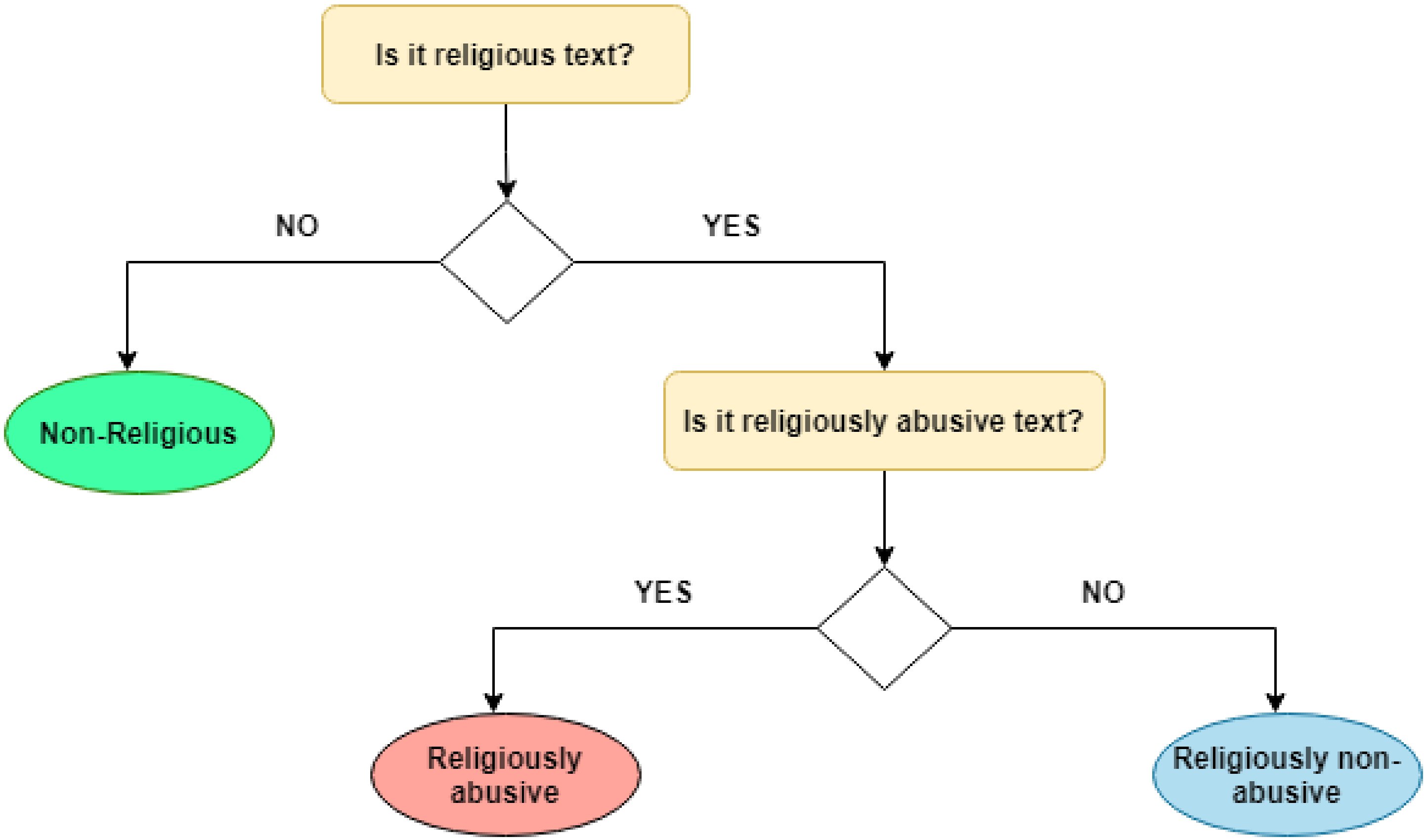

Figure 1 : Annotation Schema the corpus

Annotation Schema(Cont.)

- The labels of each category was defined manually by two annotators and then checked by a NLP expert.
- Standard rules for annotation was followed.
- The Cohen's Kappa score is calculated to determine the percentage of similarity in annotation.
- We find almost a perfect agreement on annotating the corpus.

Annotation Schema(Cont.)

	Cohen’s Kappa Value	
	Identification Level	**Classification Level**
Pair-1	0.97877	0.93502
Pair-2	0.94526	0.91524
Pair-3	0.93215	0.97112

Table 2 : Pair-wise Cohen’s Kappa values

Outline of Methodology

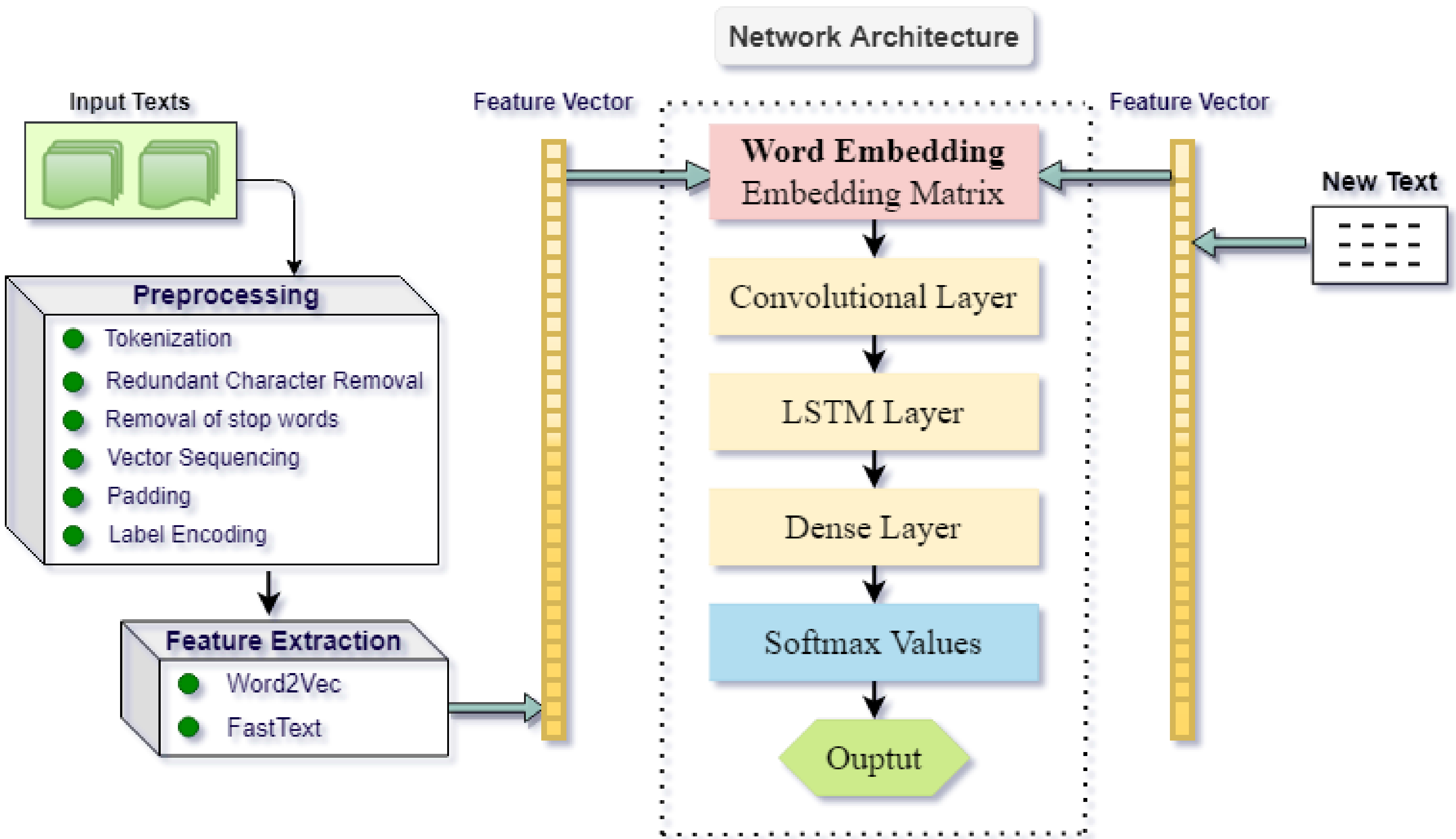

Figure 2 : Outline of the system

Outline of Methodology

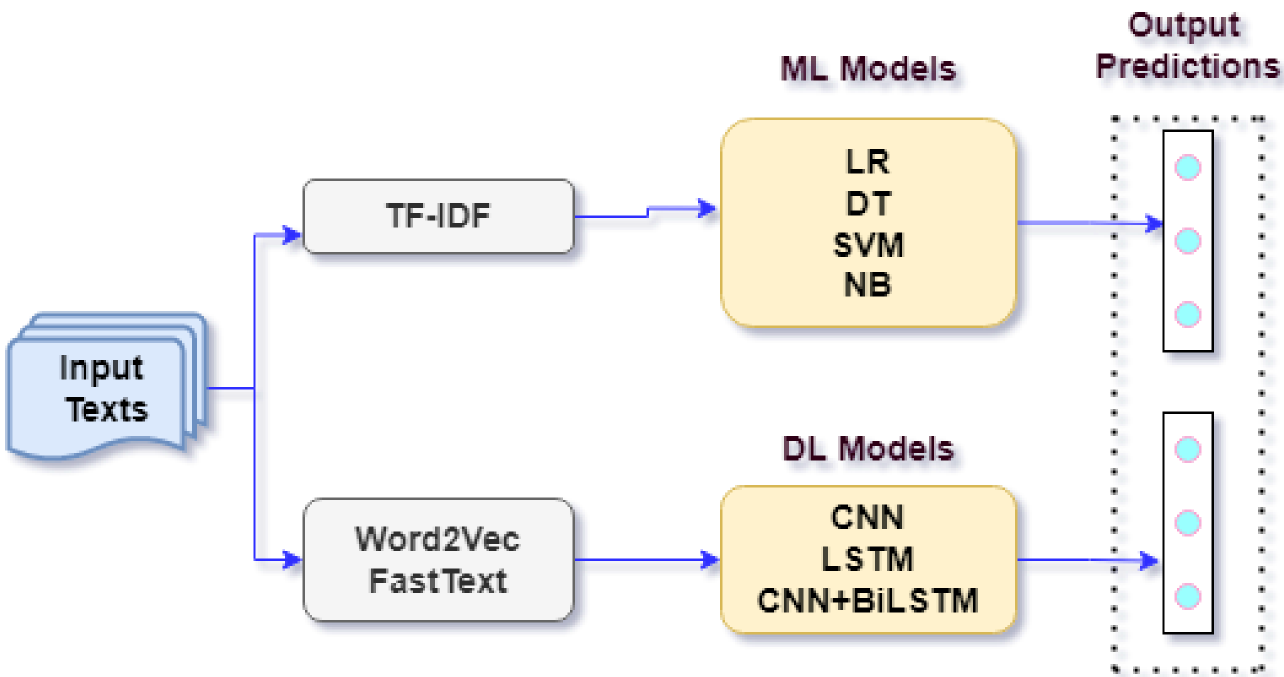

Figure 3 : Abstract Method of Religiously Abusive Comment Detection

Outline of Methodology

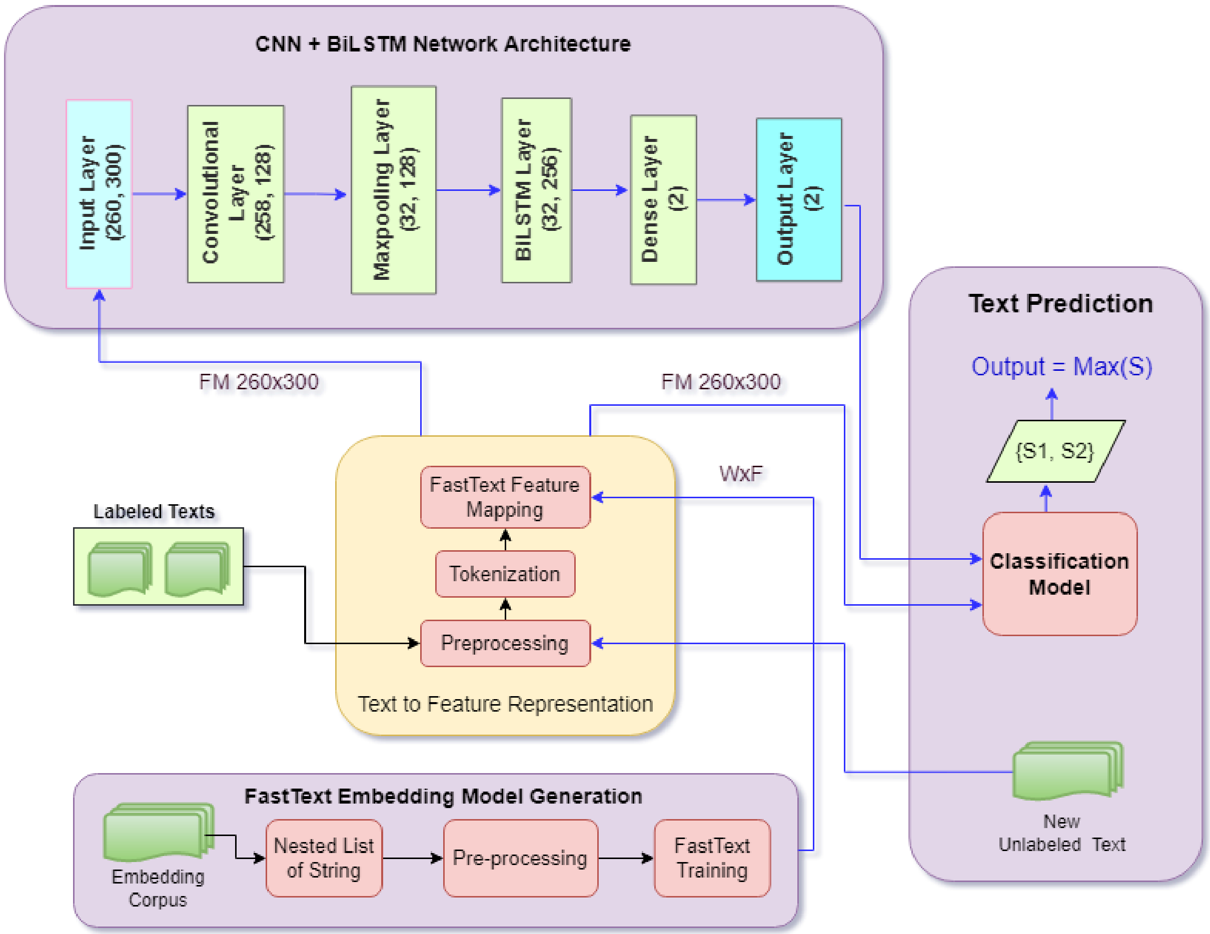

Figure 4: Framework of the System Using Deep Learning based Model

Feature Extraction

- Cleaned the dataset before extracting features.
- Extracted the textual features both for DL baselines.
- The feature vector is used to find values for both train and test
- Used pretrained FastText vectorization and embedded it through embedding layer for training DL models
  - Used embedding vector of the dimension of 100.
  - Used a pre-trained embedding matrix.

Outline of Methodology (Cont.)

Hyperparameters	Values
Filter Size	0.2
Pooling Type	'max'
Neurons in dense layer	2
Number of units	128
Batch size	32

Hyperparameters	Values
Dropout rate	0.2
Optimizer	'adam'
Learning Rate	$1e^{-5}$
Epoch	25
Batch size	32

Table 3 : List of hyperparameters values

Identification Level

		Precision	Recall	F1 score	Accuracy	AUC Area
Without FastText	**CNN**	0.79	0.59	0.60	0.59	0.6938
	LSTM	0.73	0.63	0.65	0.63	0.6641
	BiLSTM	0.80	0.76	0.77	0.76	0.7653
	CNN+BiLSTM	**0.87**	**0.86**	**0.86**	**0.86**	**0.8413**
With FastText	**CNN**	0.74	0.67	0.68	0.67	0.6776
	LSTM	0.85	0.80	0.81	0.80	0.8295
	BiLSTM	0.84	0.80	0.81	0.80	0.8222
	CNN+BiLSTM	**0.91**	**0.90**	**0.90**	**0.90**	**0.8424**

Table 4 : Comparison of Performance among the used models

Result Analysis(Cont.)

Identification Level

- Without embedding the feature vector the CNN + BiLSTM model gained the rich scores regarding of the performance parameters.

- After embedding the feature vector obtained from a pre-trained vectorizer model the performance of the combined CNN + BiLSTM model is increased more.

- The CNN + BiLSTM(FastText) performed best among all 8 classifiers.

Result Analysis(Cont.)

Classification Level

		Precision	Recall	F1 score	Accuracy	AUC Area
Without FastText	**CNN**	0.76	0.68	0.70	0.68	0.7071
	LSTM	0.73	0.57	0.59	0.57	0.6419
	BiLSTM	0.81	0.78	0.79	0.78	0.7712
	CNN+BiLSTM	**0.83**	**0.82**	**0.82**	**0.82**	**0.7903**
With FastText	**CNN**	0.72	0.70	0.70	0.70	0.6457
	LSTM	0.89	0.89	0.89	0.89	0.8658
	BiLSTM	0.90	0.89	0.89	0.89	0.8699
	CNN+BiLSTM	**0.92**	**0.90**	**0.91**	**0.90**	**0.8724**

Table 5 : Comparison of Performance among the used models

Result Analysis(Cont.)

❑ Classification Level

- Without embedding the feature vector the CNN + BiLSTM model gained the rich scores regarding of the performance parameters.

- After embedding the feature vector obtained from a pre-trained vectorizer model the performance of the combined CNN + BiLSTM model is increased more.

- The CNN + BiLSTM(FastText) performed best among all 8 classifiers.

Performance Comparison

❑ Identification Level

	Without FastText				With FastText			
	CNN	**LSTM**	**BiLSTM**	**CNN+ BiLSTM**	**CNN**	**LSTM**	**BiLSTM**	**CNN+ BiLSTM**
Religious	0.62	0.69	0.82	0.90	0.74	0.85	0.86	**0.93**
Non-Religious	0.55	0.52	0.64	0.76	0.54	0.71	0.73	**0.73**

Table 6 : Class wise performance measure in terms of f1_score

Performance Comparison(Cont.)

❑ Classification Level

	Without FastText				With FastText			
	CNN	**LSTM**	**BiLSTM**	**CNN+ BiLSTM**	**CNN**	**LSTM**	**BiLSTM**	**CNN+ BiLSTM**
Religious Abusive	0.75	0.62	0.87	0.87	0.79	0.87	0.88	**0.93**
Religious Non-abusive	0.56	0.50	0.69	0.67	0.49	0.78	0.78	**0.80**

Table 7 : Class wise performance measure in terms of f1_score

Error Analysis

- Both in Identification and Classification level the combined CNN and BiLSTM classifier outperformed the other models.

- In Identification level the CNN+BiLSTM (FastText) model gained a highest F1_score of 0.90.

- In Classification level the CNN+BiLSTM (FastText) model gained a highest F1_score of 0.93.

Error Analysis

❑ Identification Level

- The TPR (True Positive Rate) for **Religious** class is gained by the best model is 88%.
- The TPR for **Non-religious** class is obtained of 90.64%.
- The number of misclassification of Non-religious category is comparatively high.

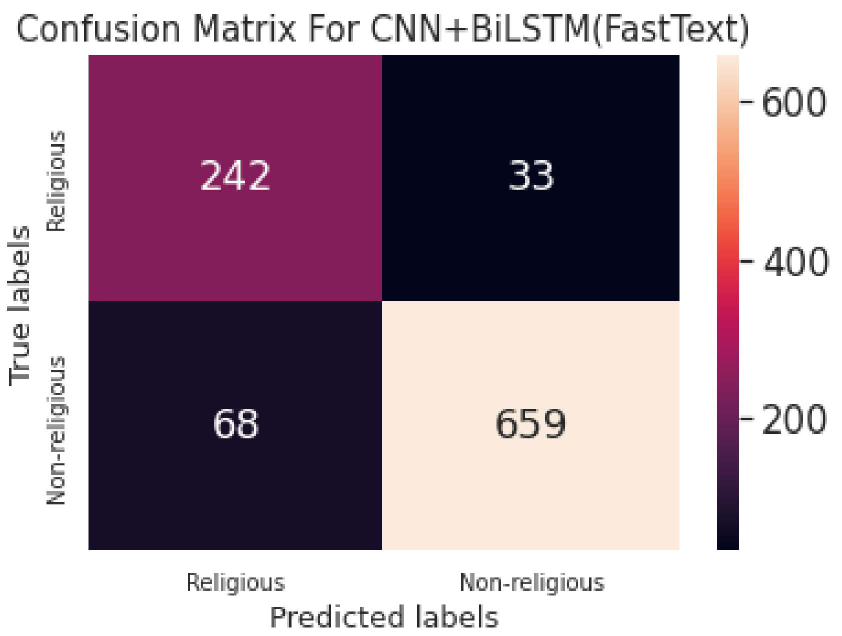

Figure 5 : Confusion Matrix for CNN+BiLSTM (FastText) model

Error Analysis

❑ Classification Level

- The TPR (True Positive Rate) for **Religious Abusive** class is gained by the best model is 80.2%.
- The TPR for **Religious Non-abusive** class is obtained of 97%.
- The number of misclassification of Non-abusive category is comparatively high.

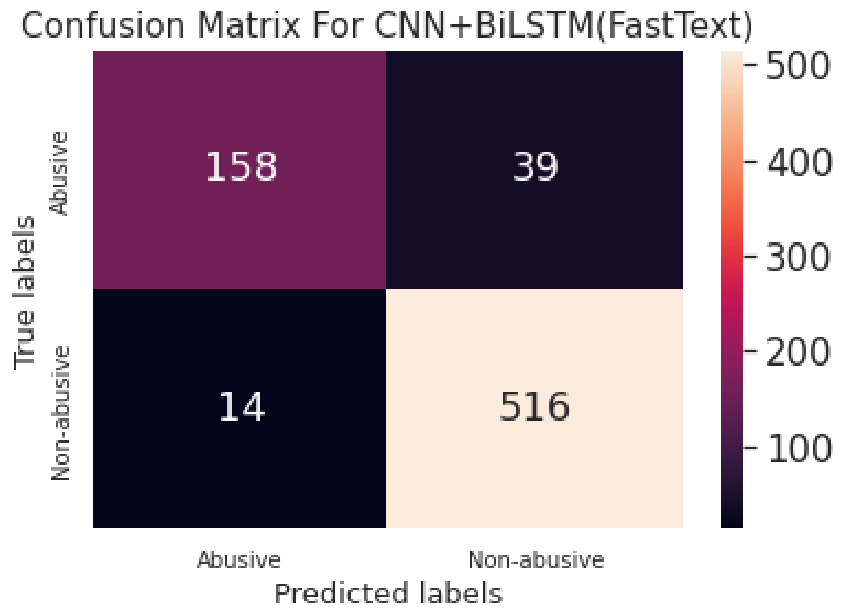

Figure 6 : Confusion Matrix for CNN+BiLSTM (FastText) model

Conclusion

- This work aims to build a corpus in Bengali using hierarchical annotation schema.

- An automated system is to be built to detect and classify religiously abusive texts in Bengali.

- An investigation must be conducted among the models' performances to choose the best model for this task.

References

1. S. Sazzed: Abusive content detection in transliterated Bengali-English social media corpus. *Online: Association for Computational Linguistics, Jun. 2021, pp. 125–130.*
2. M. O. Ibrohim and I. Budi: Multi-label hate speech and abusive language detection in Indonesian Twitter. *Italy: Association for Computational Linguistics, Aug. 2019, pp. 46-57.*
3. O. Sharif and M. Hoque: Identification and classification of textual ag- gression in social media: Resource creation and evaluation. *Apr. 2021, pp. 9–20, isbn: 978-3-030-73695-8.*
4. H.-S. Lee, H.-R. Lee, J.-U. Park and Y.-S. Han: An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems, vol. 113, pp. 22–31, 2018, issn: 0167-9236.*
5. M. Ibrohim, E. Sazany and I. Budi: Identify abusive and offensive language in indonesian twitter using deep learning approach. *Journal of Physics: Conference Series, vol. 1196, p. 012 041, Mar. 2019.*

Q & A