TOXIC COMMENT DETECTION USING BIDIRECTIONAL SEQUENCE CLASSIFIERS

MANUSCRIPT NUMBER: IDCIOT_2024 -0151

AUTHORS:

1. AMIT MAITY
2. RISHI MORE
3. PROF. ABHIJIT PATIL
4. JAY OZA
5. GITESH KAMBLI

TABLE OF CONTENTS

INTRODUCTION

- Toxic comments hurt online conversations and there is a need for systems to detect different types of toxicity like threats, insults, etc.
- The project classifies toxic comments using NLP tools (FastText, spaCy) and a BiLSTM model with two phases – evaluating toxicity and data preprocessing.
- Phase I leverages linguistic analysis and word embeddings to assess toxicity. Phase II refines the dataset for improved quality.
- Automated multi-label classification of different forms of toxicity can help to promote healthier online discussions by alerting users and enabling filtering.

LITERATURE SURVEY

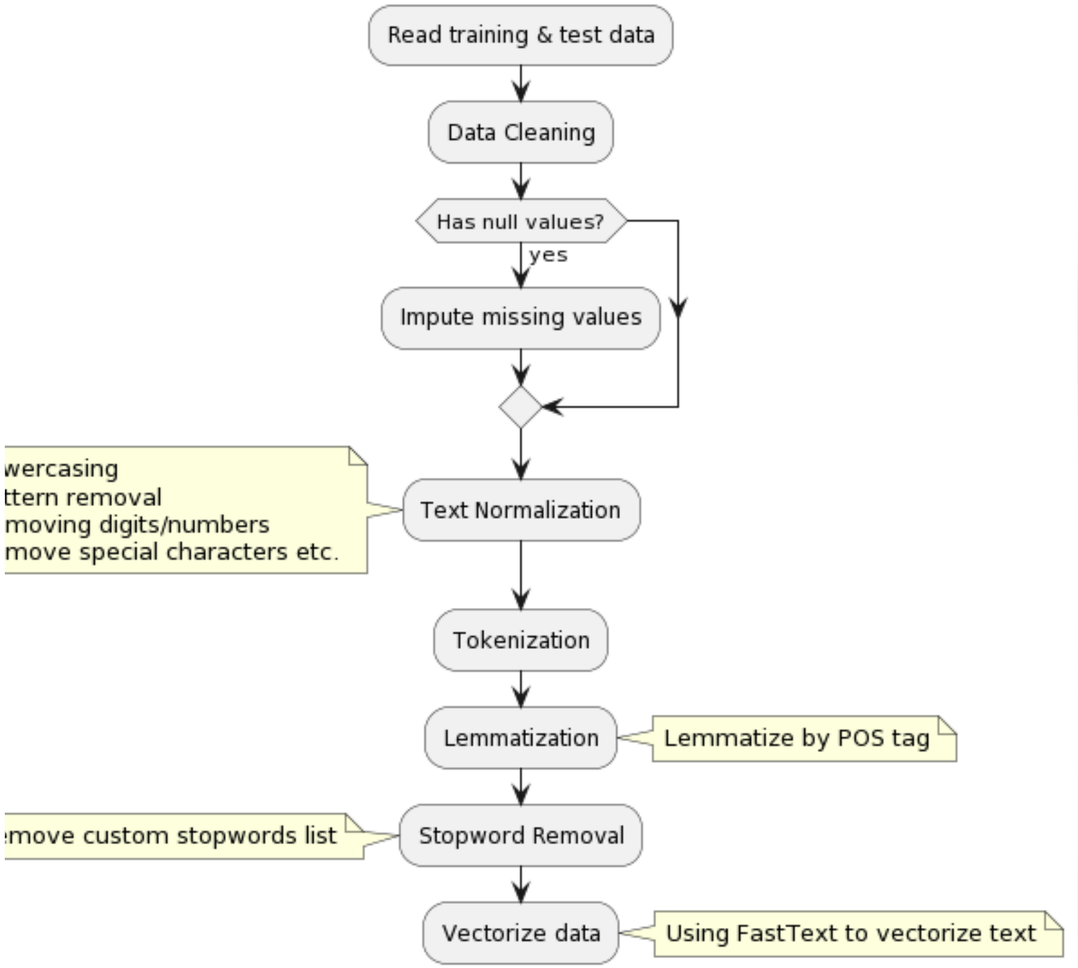
- Classical machine learning models like logistic regression, SVM, random forests have been applied for toxic comment classification, using feature engineering methods. Ensemble approaches like RVVC have achieved high accuracy.
- Deep learning models especially RNNs and LSTMs are well-suited for sequential data like text. Hybrid models combining CNNs and LSTMs have been proposed that learn both local and global patterns.
- Research has explored detecting toxicity in diverse contexts beyond comments, like song lyrics. Multi-label classification has also been studied for categories like threats, insults.

LITERATURE SURVEY

- State-of-the-art models utilize word embeddings and transfer learning with models like XLMRoBERTa. These achieve strong results across multilingual datasets.
- BiLSTM's memory capabilities are optimal for retaining context in comments. Attentional, multi-channel BiLSTM networks with rich word representations have achieved new benchmarks in hate speech detection.

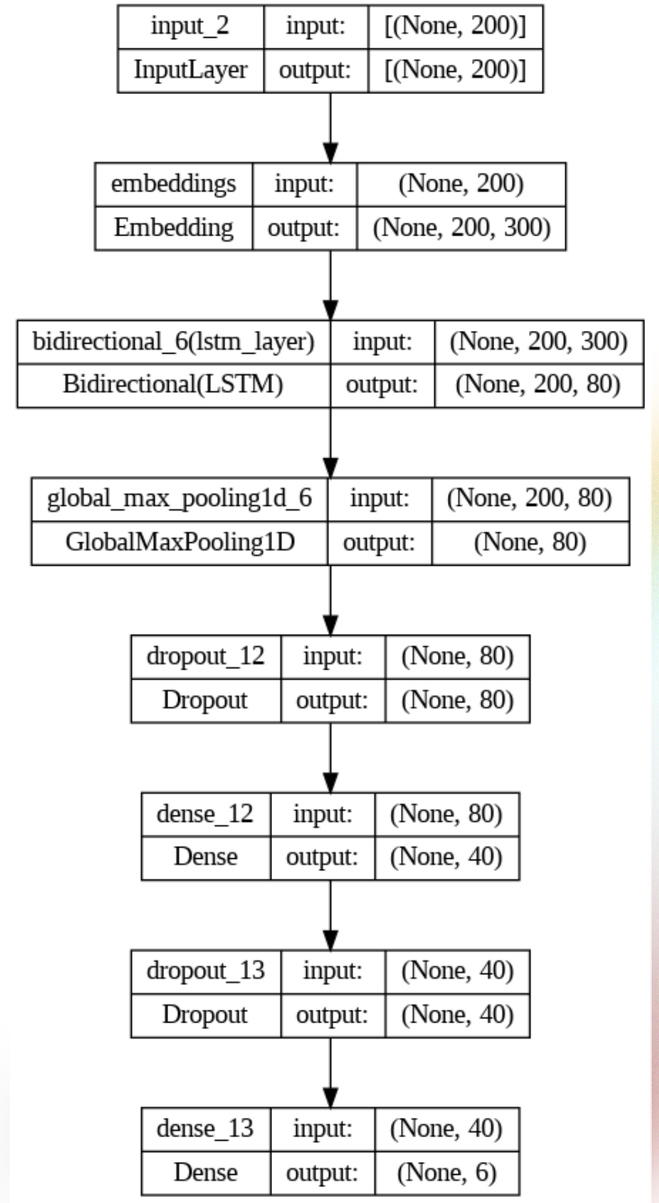
METHODOLOGY

- The dataset from Kaggle has 159k rows labeled as toxic/non-toxic, insults, identity hate etc. EDA provides insights like 20% comments being toxic.
- Text preprocessing steps include lowercasing, pattern removal, lemmatization, stopword removal etc. to handle noise in social media data.
- Tokenization and padding ensure fixed length input sequences for the model. FastText embeddings capture semantics.

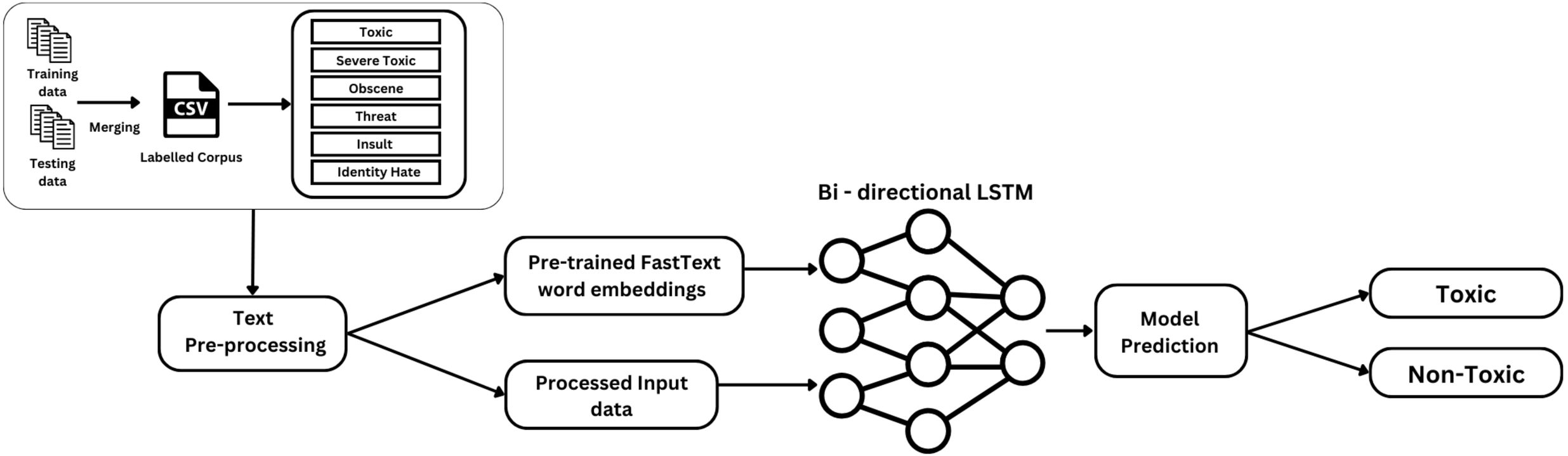


METHODOLOGY

- A BiLSTM neural network architecture is proposed with embedding, BiLSTM, max pooling, dropout and dense layers.
- The model is trained to minimize a loss function via backpropagation. Hyperparameters are tuned using validation set.
- Evaluation involves test set and metrics like accuracy, precision, recall. Confusion matrices give further insights.



SYSTEM ARCHITECTURE



RESULTS

- The model performs well on all four toxicity types, with F1-scores above 0.85 for all categories.
- The model is particularly good at identifying "Insult" examples, with a precision of 0.96 and a recall of 0.97.
- However, it is worth noting that the model's performance may be different on different datasets, and it is important to carefully evaluate the model's performance on the specific task for which it is being used.

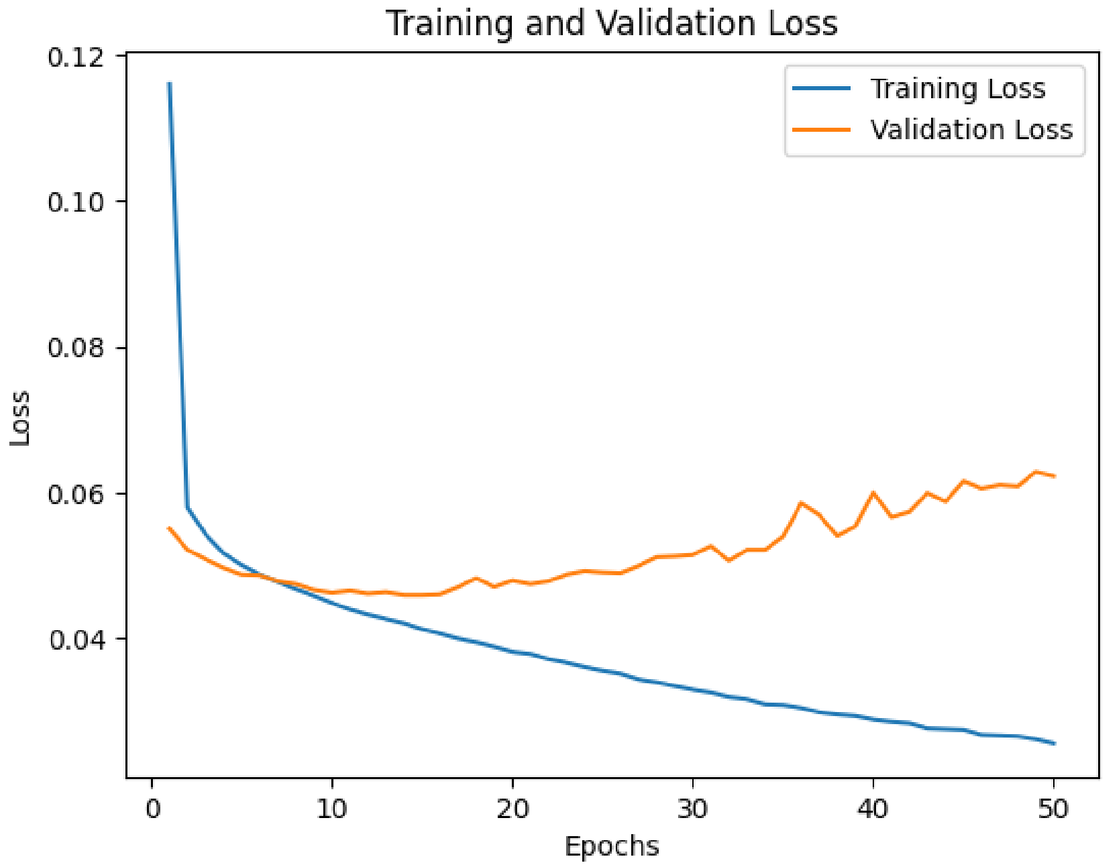
TABLE I
TRAINING AND VALIDATION LOSS AND ACCURACY AT SELECT EPOCHS.

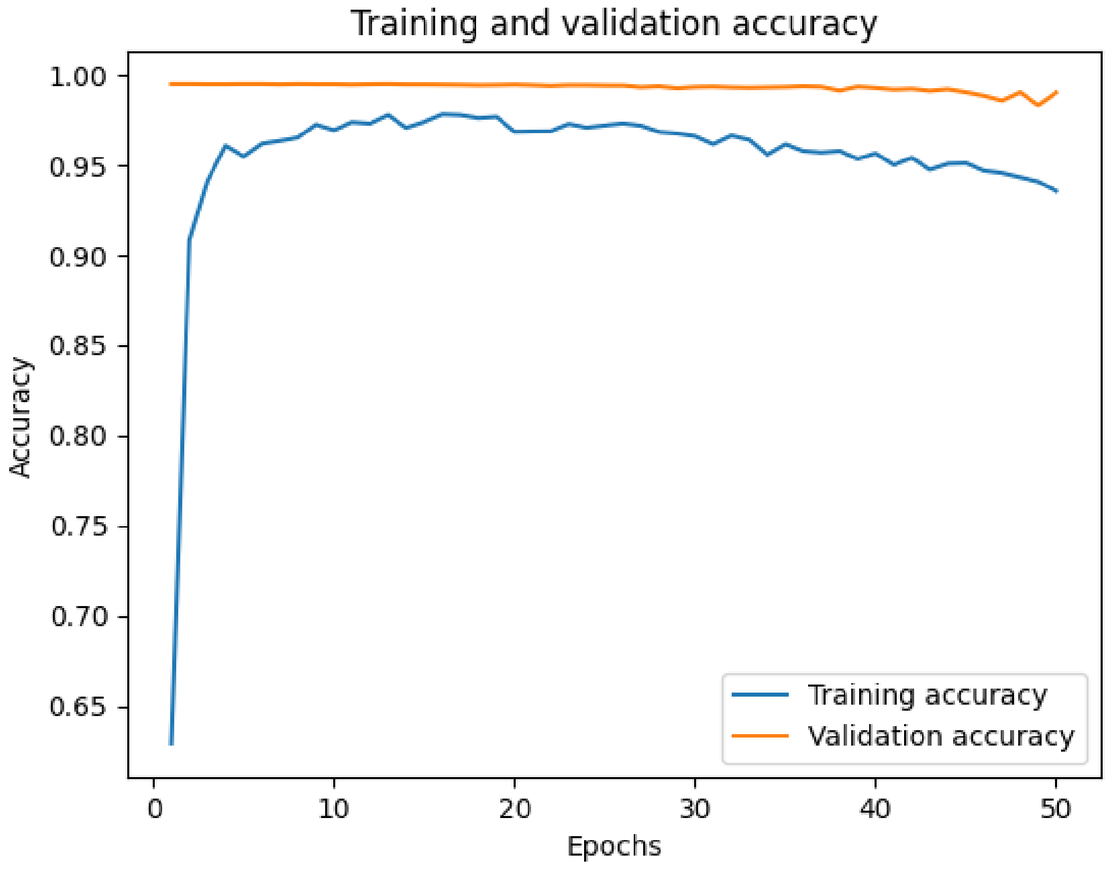
Epoch	Train Loss	Train Accuracy	Val Loss	Val Accuracy
1	0.116	0.629	0.055	0.994
10	0.0448	0.969	0.0462	0.994
25	0.0355	0.972	0.049	0.994
50	0.0255	0.935	0.0622	0.99

TABLE II
PERFORMANCE METRICS OVER 100 EPOCHS

Toxicity Type	Precision	Recall	F1-Score
Toxic	0.94	0.93	0.93
Severe Toxic	0.91	0.89	0.90
Obscene	0.93	0.95	0.94
Threat	0.87	0.85	0.86
Insult	0.96	0.97	0.96
Identity Hate	0.89	0.88	0.88

RESULTS


Training and Validation Loss
Training Loss
Validation Loss
Loss
Epochs
0.12
0.10
0.08
0.06
0.04
0
10
20
30
40
50


Training and validation accuracy
Training accuracy
Validation accuracy
Accuracy
Epochs
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0
10
20
30
40
50

CONCLUSION

- The study proposes an optimized BiLSTM neural network architecture that achieves over 95% accuracy in multi-label toxic comment classification outperforming prior works.
- Key contributions include updated datasets, optimized architecture for multi-label classification, state-of-the-art accuracy, and a deployable solution for large-scale filtering.
- The work represents significant progress in using AI to detect diverse forms of textual toxicity, enabling the mitigation of harmful content for social good.

FUTURE SCOPE

- Integrate large pre-trained language models like BERT and RoBERTa through fine-tuning to enhance semantic understanding and generalization.
- Expand training data diversity and size using augmentation techniques like backtranslation and multi-task learning for greater robustness.
- Build ensemble models combining complementary neural architectures like LSTMs, CNNs and transformers to improve accuracy.

THANK YOU