Emotion Analysis and Classification using a Hybrid CNN-BiLSTM approach

Shreenika Aldur Krishnegowda
230016400
Msc (Data Science)
Shreenika.Aldur-Krishnegowda@city.ac.uk

Google-Colab-Link:
https://drive.google.com/drive/folders/1S0aB6dwOUYghERsB-J3ysRDRhAO4mDjS

1 Problem statement and Motivation

Emotions are the prime aspects of human beings and the ability to freely express them and act upon them is one of the fundamentals of being a human. They are critical to human interaction and one of the primary aims of developing smart technologies such as AI has always been to understand and respond to human needs.

Social media is huge playing field where millions gather and express their sentiments in the form of textual data. It generates an abundance of information that can be leveraged to understand and analyze individual emotions and public sentiments. However, emotion classification remains a problem due to the complexities in human language.

This paper aims at compounding deep learning models that can generalize across diverse linguistic expressions and contexts while offering higher precision and maintaining high performance. This paper aims to study the application of Convolutional Neural Network (CNN) architecture and a new hybrid CNN-BiDirectioanl Long Short-Term Memory (CNN-BiLSTM) model on the emotion dataset obtained from Hugging Face. It starts with developing an understanding of the dataset and its nature then divulges into proving the the proposed hypothesis. This is done with implementing a new approach of a hybrid model with CNN-BiLSTM and then treating the model with Hyperparameter optimization using a combination of Adam and Bayesian optimisation.

2 Research hypothesis

This paper is based on two major hypotheses:
Hypotheses-A: Due to its enhanced feature extraction capabilities the CNN-BiLSTM model will outperform the standalone CNN model in the classification of emotions.

Hypothesis-B: LSTM's capacity to process sequences will reduce the need for extensive feature engineering thus allowing CNN-BiLSTM to achieve equal or better performance than standalone CNN model with a same number of epochs.

Hypothesis-A is based on the idea that since CNNs are adept at extracting spatial features, it can identify patterns in word usage and sentence structure. The convolutional filters will then act as feature detectors for emotions embedded in texts. LSTMs on the other hand excel at capturing temporal relationships in sequential data allowing them to interpret the sentiment and emotion of the target sentence.

Hypothesis-B is based on the fact that LSTMs reduce the need for manual feature engineering since they can automatically learn to recognize and utilize temporal structure in the data. They retain information for a longer period of time allowing them to use context from previous texts to inform later predictions. This ability will allow the hybrid CNN-BiLSTM model to converge faster during training compared to the standalone CNN model.

3 Related work and background

Emotion classification and sentimental analysis gained its popularity post 2004 afer the booming of social media which result in 99% of all the research papers being published post 2004. In recent years the focus of emotion analysis and sentiment analysis has shifted from product reviews to social media platforms like Twitter (now X) and Facebook (now Meta). The research further led to the Sentiment Analysis process in the paper "A review on sentiment analysis and emotion detection from text" by (Nandwani, P., Verma, R.). The authors exclaimed that in the categorical model, emotions are defined discretely, such as anger, happiness, sadness, and fear. Depending upon the particular categorical model, emotions are categorized into four, six, or eight categories (Nandwani, P., Verma, R.). Hence, providing the motivation behind selecting the dataset used in this paper. For emotion classification they adopted feature engineering to enhance their deep neural models.

The next research paper reviewed was (Ramadhani and Goo) which conducted Analysis on 4000 Korean and English tweets. They used Deep Neural Network with three hidden layers and the model was optimized using Stochastic Gradient Descent (SGD). The model managed to achieve a 75.03% accuracy on the dataset. Similarly (Dholpuri et al. 2018) conducted Sentiment Analysis on IMDB dataset and compared a variety of machine learning and deep learning models. The text from dataset was processed to remove irrevalant characters and repeating words and naïve bayes, SVM, logistic regression, k-nearest neighbor, and a CNN was applied. The study showed that CNN achieved 99.3% accuracy, the highest of all the proposed models.

Further going down, (Yang, Y. 2018) in their paper Convolutional Neural Networks with recurrent neural filters, proposes a RNN-filter based CNN and LSTM model to approach Sentiment Analysis. They have used Stanford Sentiment Treebank Dataset and the model comprised on a pooling layer and an LSTM layer. Adam Optimizer and early stopping was adopted to avoid overfitting. The model managed to achieve 53.4% accuracy on the dataset. Similarly, (Goularas and Kamis 2019) proposes a hybrid model combining the strength of CNN and LSTM. They used tweets dataset obtained from International Workshop on Semantic Evaluation competition, the authors conducted various preprocessing steps as well as converted text to lower characters. One distinction to be observed was the use of word2vec and GloVe word embeddings. The Hybrid CNN and LSTM model with GloVe embeddings managed to achieve 59% accuracy.

Building up on the idea of a Hybrid CNN-LSTM model, the study by (Rhanoul et al. 2019) and (Tyagi et al. 2020) both presented the idea of combining a CNN with bidirectional Long Short-term Memory (BiLSTM) network for their analysis. (Rhanoul et al. 2019) experimented on dataset containing 2003 articles and news articles, text was processed using word embedding with pretrained doc2vec model. The proposed model by the study consisted of convolutional layer, a max-pooling layer, a dropout layer and BiLSTM layer and managed to achieve an accuracy of 90.66% on the dataset. (Tyagi et al. 2020) operates on Sentiment140 dataset consisting of 1.6 million tweets and used preprocessing to clean the data. Similarly, (Tyagi et al. 2020) also utilized an embedding layer but with GloVe pretrained model and utilised convolutional layer with a BiLSTM layer and dropout layers. The model achieved an accuracy of 81.20% accuracy on the dataset. Interestingly, both the (Rhanoul et al. 2019) and (Tyagi et al. 2020) papers apply the hybrid CNN-LSTM model on positive, negative and neutral sentiments instead of more complex emotions.

(Chundi et al. 2020) also developed a Hybrid CNN-LSTM model but experimented on a dataset containing 10,401 comments in English, Kannada, and a mixure of both languages. The hybrid model scored an effective accuracy of 77.6% on the multilingual dataset. (Jang et al. 2020) proposed a Hybrid CNN and LSTM model for the IMDB dataset. It contains 50k positive and negative sentiments and used word2vec pretrained embedding model for preprocessing and implemented Adam optimizer and dropout to reduce overfitting. The model achieved 90.26% accuracy on the IMDB dataset for sentiment analysis.

The literature review sheds light on the implementation of Hybrid CNN-BiLSTM model justifying the approach proposed in this paper. Where the paper deviates from the above studies conducted is the usage of hyperparameter tuning using the combination of Adam and Bayesian optimization techniques on a multiclass classification dataset with varied emotion labels (more details and discussion in the Section 6: Dataset) and testing the proposed hypothesis.

4 Accomplishments

1. Task 1: Loading the emotion dataset from hugging face and defining the dataset structure. The dataset is pre-split into Train, Test and Validation sets. Applying Exploratory Data Analysis to understand the distribution of each emotion in train set and statistical analysis to get average word count - Accomplished.

2. Task 2: Data preprocessing task of cleaning the data (remove special characters, whitespace, convert to lower case). – Accomplished.

3. Task 3: Tokenising the word text data from Train and validation sets. – Accomplished.

4. Task 4: Padding the sequences of the texts from train and validation sets – Accomplished.

5. Task 5: Implementing the standalone CNN model and generating Train and Validation loss per epoch and saving the Deep learning model – Accomplished.

6. Task 6: Implementing the Hybrid CNN-BiLSTM model and generating train and Validation loss per epoch and saving the Deep learning model – Accomplished

7. Task 7: Treating the Hybrid CNN-BiLSTM model to Hyperparameter tuning using a combination of Adam and Bayesian Optimization – Accomplished

8. Task 8: Evaluating the three models on Test data and choosing the best models – Accomplished

9. Task 9: Generating comparison matrix for all the models for critical evaluation – Accomplished

10. Task 10: performing final analysis on the Hyperparameter tuned optimized Hybrid CNN-BiLSTM model - Accomplished

5 Approach and Methodology

Convolutional Neural Network: They are a class of deep learning neural network primarily used for classfication tasks. They are designed to automatically adapt and learn spatial hierarchies of features directly from the textual input data (Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012). While training the labelled dataset of text, CNNs can automatically learn to extract relevant features from text and map them to appropriate emotions.

BiDirectional Long Short-Term Memory Networks: They are a type of recurrent neural network architechure mostly used in sequence model tasks such as emotional analysis on text data. BiLSTMs are known for their ability to capture both the past and future context information. LSTM is the building block BiLSTMs, it processes the input sequence in two directions. One set of LSTM cells will process the input sequence from beginning to the end and the other one from end to the beginning. This approach allows the network to capture both past and future context information from each token of the sequence (schusters, M., & Paliwal, K. K. 1997).

Adam Optimizer: Adaptive Moment Estimation is a popular optimization algorithm used in the training of deep learning networks such as CNN and RNN models.

Bayesian Optimization: This is a powerful technique used to search and find the best input parameters to maximize or minimize the model's output. It is less expensive than other optimization techniques which have high computational costs. This approach makes it well suitable for emotion classification from text data using complicated approach such as CNN and BiLSTM.

The approach this research ended up taking involved training three different models for emotion classification. As a baseline approach a simple CNN model was trained then a Hybrid CNN-BiLSTM model and finally a Hyperparameter-Tuned Hybrid CNN-BiLSTM Model. Each of these models leverages upon different techniques to classify emotions. The core idea behind this approach is leveraging upon the strength of convolutional neural network and recurrent neural network specifically the Bidirectional Long Short-Term Memory layers. BiLSTMs are noted for their ability to capture sequential information from text.

Data Preprocessing involves cleaning the text data, tokenization and padding the features for uniformity. The effectiveness of all these models were then compared to each other and it was oberved that the Hybrid model performed significantly better then the baseline CNN model however, the hyperparameter tuned Hybrid CNN-BiLSTM model gave a more optimized performance.

The proposed approach in this paper introduces a more complex CNN architecture and hyperparameter tuning. However, the approach still faced the similar limitations as the baseline model. The models reliance on the quality and quantity of training data may still lead to overfitting, and the unbalanced structure of the data produced miscalculations although negligible.

6 Dataset

The dataset utilised in this paper is acquired from Hugging Face public website: https://huggingface.co/datasets/dair-ai/emotion. The dataset was found in .jasonl format and is pre-split into Train, Test and Validation sets. The train, test and validation split contained 2 sub-categories namely text and label. Text contains string feature and label is a classification field consisting of 6 [0,1,2,3,4,5] values. Each value in label indicates a target emotions: 0 – sadness, 1 – joy, 2 – love, 3 – anger, 4 – fear, 5 – surprise. The dataset was directly imported into the notebook using “`emotion_dataset = load_dataset("dair-ai/emotion")`” command.

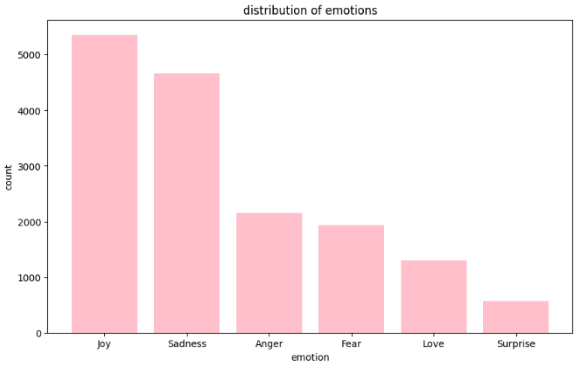
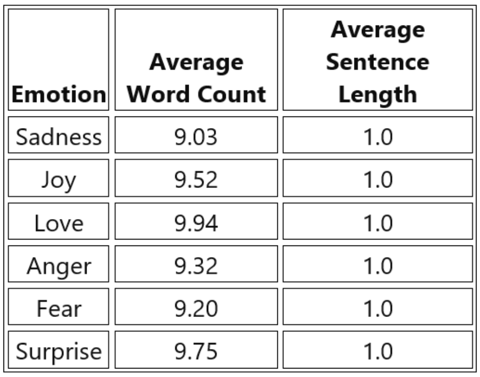


Figure 1: Frequency distribution of each classes

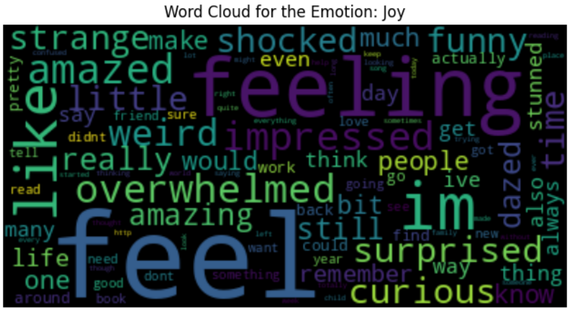
As per Figure 1, its evident from performing initial analysis of the dataset that there exists slight imbalance between target classes. The Emotion of Joy exceeds 5000 values and the lowest emotion of surprise has under 1000 values. For the purpose of this paper it was decided to keep this as is.

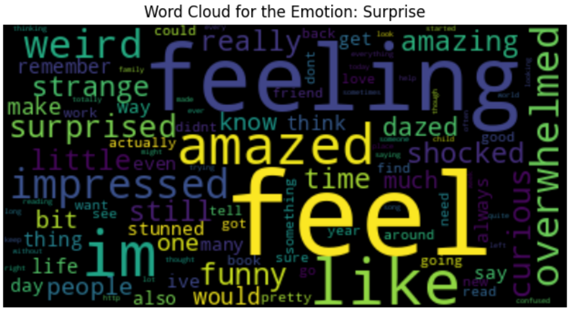


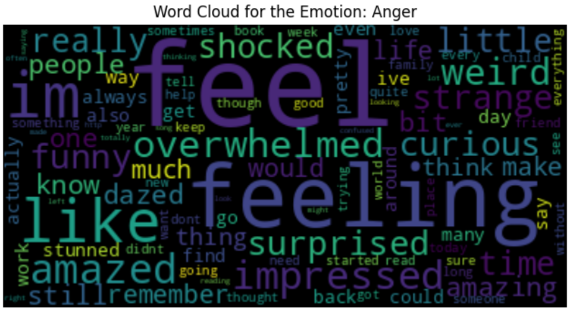
Emotion	Average Word Count	Average Sentence Length
Sadness	9.03	1.0
Joy	9.52	1.0
Love	9.94	1.0
Anger	9.32	1.0
Fear	9.20	1.0
Surprise	9.75	1.0

Figure 2: Average word count and sentence length

The average word count across all the emotions is quite similar. Ranging from 9 to 10 words per text in the dataset. The average sentence length is consistent across all the emotions in the dataset with each sentence approximately containing 1 period, indicating that all the text in the train set consists of short simple sentences.







These word clouds formed using BoW from the joy, surprise and anger labelled classes shows the token distribution for the highly misclassified labels by the baseline CNN model.

6.1 Dataset preprocessing

Data processing was carried out in 2 major phases namely: a) Cleaning of the train, test and validation subcategories and b) Tokenizing the cleaned data and preparing them for model training.

a) Cleaning the data: This step ensure that the text data cleaned and processed properly before using it more model training. The steps carried out during this phase involved removing the duplicate values present in the dataset, splitting of the train, test and validation dataframes into features (X) and labels (y), with X representing the text data and y representing their corresponding emotions. Finally, cleaning the data by removing special characters, converting it lower case, then splitting the text into tokens, removing the stop words and finally lemmatizing the tokens to reduce them to their base form.

b) Tokenizing the cleaned data and preparing them for model training: To begin with, the cleaning features (X) and labels (y) of train and validation data frames are converted to numpy arrays. For data preprocessing it was decided to initialize a tokenizer with maximum vocabulary size of 5000 words. Only the 5000 most frequent words in training data will be considered for tokenization and finally the tokenizer was fit on training data. Finally, the features of both train and validation data frames were converted to a sequence of integers ready for model training.

7 Baselines

This paper is built on a straightforward CNN architecture which is easy to interpret and implement. It serves as a model for the more complex architectures implemented in the later stages. The baseline model has various limitations and the accuracy suffers in the final evaluation. This baseline aligns with the idea of emotional classification from textual data. The baseline model is build on the same data preprocessing steps and tokenization techniques which is used for hybrid-CNN-BiLSTM model and hyperparameter tuned CNN-BiLSTM model. The baseline model achieves a reasonable performance while training but while evaluating the trained model the performance sharply decreases, serving as a starting point for the study to improve upon the performance and tackling the limitations in its implementation.

8 Results, error analysis

The final analysis of the paper is divided into two phase training and testing .

Training Phase:

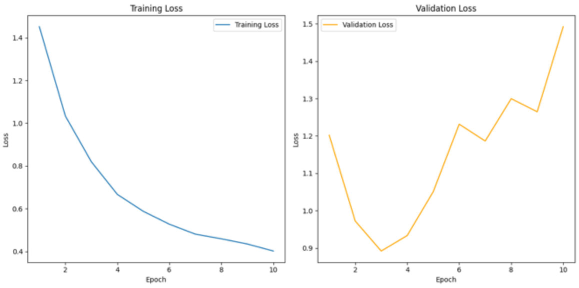


Figure 3: Train and Validation loss versus epoch for baseline model.

In the baseline model the training loss gradually decreases over epoch indicating that the model is learning from the training data ,however the validation loss shows some fluctuations and increases around the third epoch. the model doesn't perform optimally on the unseen data.

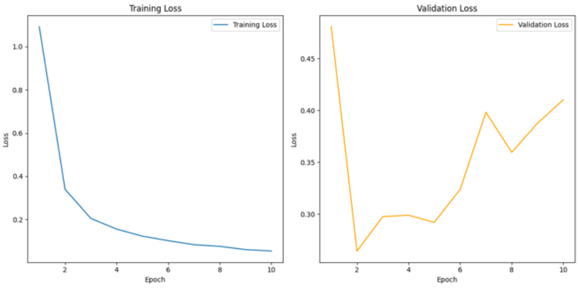


Figure 4: Train and Validation loss versus epoch for hybrid CNN-BiLSTM model.

Hybrid CNN-BiLSTM model shows a lower validation loss compared to the baseline model. But validation loss after the third epoch increases.

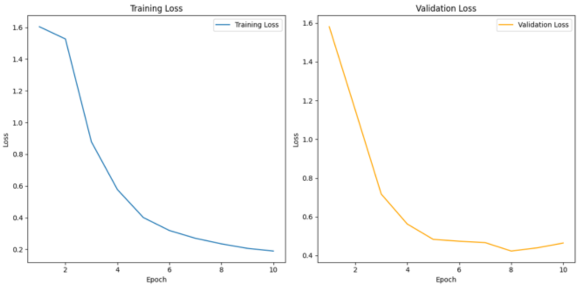
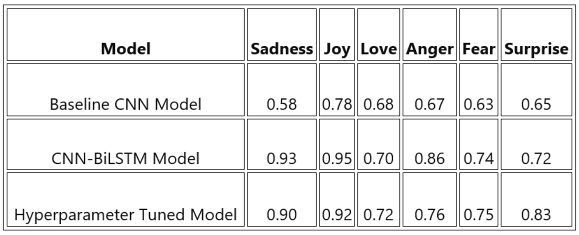


Figure 5: Train and Validation loss versus epoch for hyperparameter tuned CNN-BiLSTM model.

Hyper parameter tuned model shows the best performance and the validation loss and the validation loss doesn't increase at the layer epoch, it better generalizes the unseen data than other models.



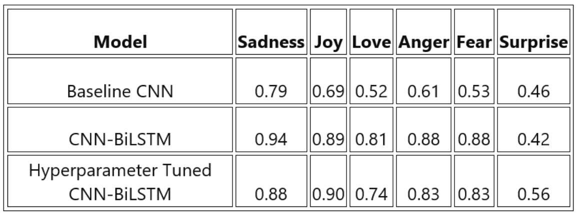
Model	Sadness	Joy	Love	Anger	Fear	Surprise
Baseline CNN Model	0.58	0.78	0.68	0.67	0.63	0.65
CNN-BiLSTM Model	0.93	0.95	0.70	0.86	0.74	0.72
Hyperparameter Tuned Model	0.90	0.92	0.72	0.76	0.75	0.83

Table 1: Precision per class of every model

The baseline model achieved reasonable precision ranging from 0.58 to 0.78, proving to be a reasonable and decent baseline for the emotion classification and analysis. However, the baseline model has room for improvement in the precision of emotion labels. The model struggles in predicting sadness, fear and surprise but achieves a higher precision for joy, indicating room for improvement in capturing other emotions.

This problem was addressed by incorporating a new approach that is the Hybrid CNN-BiLSTM model which managed to achieve a significantly higher precision rate of 0.70 to 0.95. This model improves on the baseline model in every class especially showing higher precisions in sadness which the baseline model struggles the most. However, the Hybrid CNN-BiLSTM Model still exhibits slight lower performance in predicting love and surprise.

Finally, the hyperparameter tuned CNN-BiLSTM model demonstrates consistent improvements in precision across all the emotion classes compared to the other two models. Especially in predicting love and surprise emotions on which both the baseline and hybrid CNN-BiLSTM model achieved poor performance, indicating the effectiveness of hyperparameter tuning for optimizing the model performance.



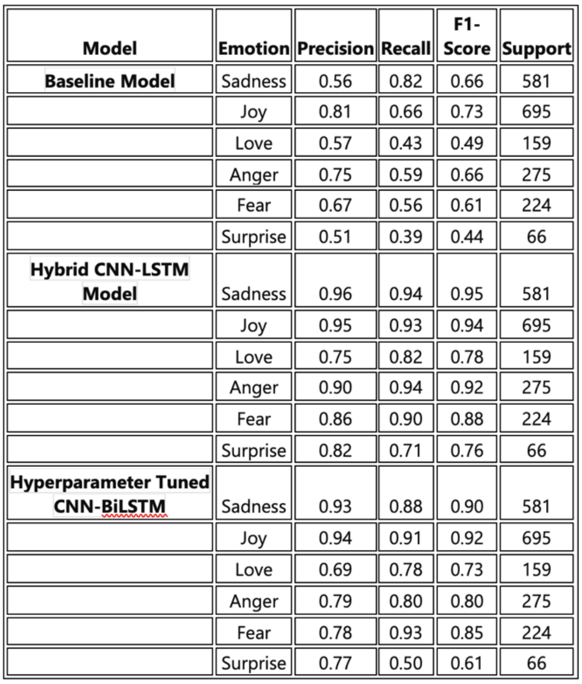
Model	Sadness	Joy	Love	Anger	Fear	Surprise
Baseline CNN	0.79	0.69	0.52	0.61	0.53	0.46
CNN-BiLSTM	0.94	0.89	0.81	0.88	0.88	0.42
Hyperparameter Tuned CNN-BiLSTM	0.88	0.90	0.74	0.83	0.83	0.56

Table 2: Recall scores per class of every model

The baseline model posts moderate recall scores for most emotions showing higher performance in sadness but similar to precision struggles in emotions like love and surprise showing low recall score of 0.52 and 0.46 respectively. The Hybrid CNN-BiLSTM model remarkably high recall scores in comparison to the baseline model but still struggles in emotions like love and surprise. For surprise emotion the recall score drops lower than the baseline model which caused significant problems in the model approach.

Finally, the hyperparameter tuned CNN-BiLSTM model managed to solve this issues as it leveraged on BayesianOptimization and keras-tuner to iterate through all the hyperparameters and finding the best one to train the model. The recall score for surprise notably improved compared to both the baseline and hybrid CNN-BiLSTM model.

Testing Phase:



Model	Emotion	Precision	Recall	F1-Score	Support
Baseline Model	Sadness	0.56	0.82	0.66	581
	Joy	0.81	0.66	0.73	695
	Love	0.57	0.43	0.49	159
	Anger	0.75	0.59	0.66	275
	Fear	0.67	0.56	0.61	224
	Surprise	0.51	0.39	0.44	66
Hybrid CNN-LSTM Model	Sadness	0.96	0.94	0.95	581
	Joy	0.95	0.93	0.94	695
	Love	0.75	0.82	0.78	159
	Anger	0.90	0.94	0.92	275
	Fear	0.86	0.90	0.88	224
	Surprise	0.82	0.71	0.76	66
Hyperparameter Tuned CNN-BiLSTM	Sadness	0.93	0.88	0.90	581
	Joy	0.94	0.91	0.92	695
	Love	0.69	0.78	0.73	159
	Anger	0.79	0.80	0.80	275
	Fear	0.78	0.93	0.85	224
	Surprise	0.77	0.50	0.61	66

Table 3: Classification report of 3 models

During the testing phase the baseline mode, the hybrid CNN-BiLSTM model and the hyperparameter tuned cnn-bilstm model was tested against the test dataset. The results does not deviate much from the training phase.

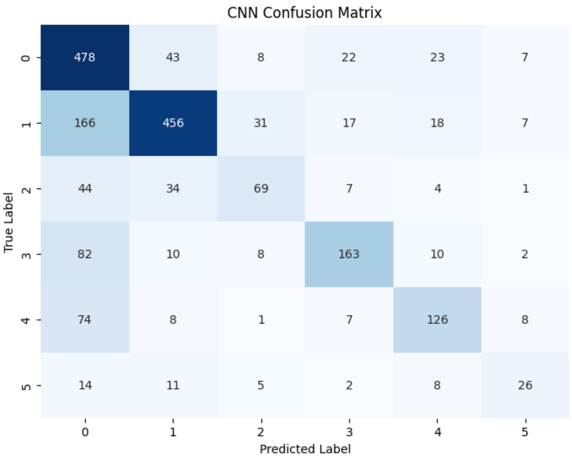


Figure 3: Confusion Matrix for baseline model

The baseline CNN model achieves a strong overall accuracy of 66%, with its strength being performing well in predicting joy emotion label with 81% precision and 66% recall score. However, the model struggles in predicting love and surprise label.

Error analysis:

The baseline Model as shown in Figure 3 misclassifies instances of joy as sadness and vice-versa and also struggles to distinguish between anger and fear labels. This was identified to be caused by the similarity in tokens of both these class labels. As evident by figure which shows the wordcloud formed by the BoW of joy, surprise and anger label, most of the tokens used to classify these classes are repeated. This causes ambiguity in classification as certain tokens are highly indicative of certain emotions while other tokens are simply ambiguous. To fix this tokenization was implemented using keras tokenizer however, the misclassification still persisted between these class labels. It could be fixed using feature engineering however since the aim of this paper is to acheive emotion classification and model performance without relying on feature engineering we move towards the proposed hybrid CNN-BiLSTM model leveraging on the sequential and spatial capabilites of BiLSTM and CNN architecture as well as hyperparamter tuning of the hybrid cnn-bilstm model to address the issue.

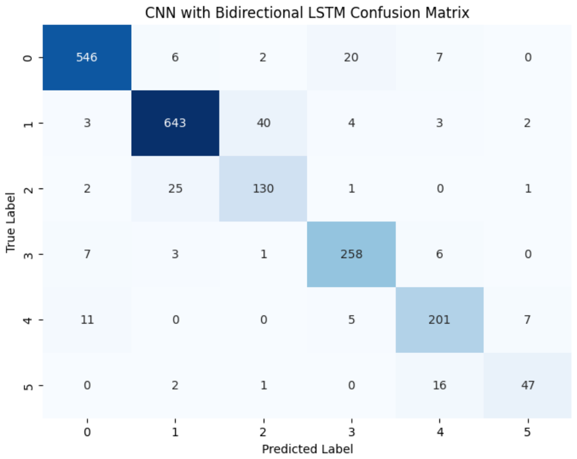


Figure 4: Confusion matrix for Hybrid CNN-BiLSTM model

The hybrid CNN-BiLSTM model shows excellent performance and improvement over the baseline model by demonstrating an over accuracy of 91% with its strength being predicting sadness labels and a strong performance in anger and fear label. However, the model occasionally misclassifies the instances of surprise label as evident by the confusion matrix.

The Hybrid CNN-BiLSTM model demonstrated improved performance compared to the baseline CNN model. As evident by figure 4, the hybrid model shows fewer misclassifications then the baseline CNN model.

Error Analysis:

The Hybrid CNN-BiLSTM model exhibits improved classification compared to the baseline model and although some misclassification can be observed between joy and anger it is a significant improvement over the baseline model.

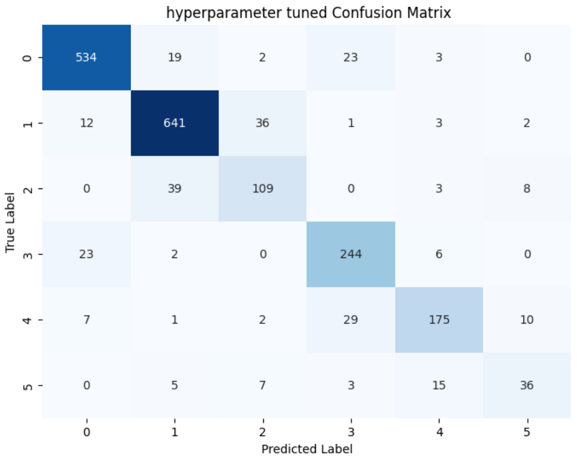


Figure 5: Hyperparameter tuned CNN-BiLSTM model

The Hyperparameter tuned CNN-BiLSTM model achieves a moderate accuracy of 87% showing similar strength to the hybrid model and performs strongly in sadness and joy class labels demonstrating higher precision.

Error Analysis:

The hyperparameter tuned CNN-BiLSTM model demonstrates similar classification competency as the hybrid CNN-BiLSTM model if anything the model appears to be more consistent with its classification.

9 Lessons learned and conclusions

This paper manages to achieve its goal of successfully improving upon the baseline model and proving the hypothesis stated by it. Moreover with the implementation of hyperparameter tuning the study further improved upon the performance of the hybrid CNN-BiLSTM model and managed to secure consistent performance for emotion analysis. Though we observe that the accuracy decreases in the hyperparameter tuned CNN-BiLSTM model the performance remains consitstent and loss between training and validation remains consistent unlike the baseline CNN and hybrid CNN-BiLSTM model.

This study heavily relies on the unbalanced nature of the dataset and strengths of both CNN and LSTM. Leveraging on the fact that CNN can effectively capture spatial features and BiLSTM can capture sequential information, we were able test the trained models without relying on feature extraction methods proving the effectiveness of this approach. The hyperparameter-tuned CNN BiLSTM model further displayed consistent performance and improved on shortcomings of both the baseline and the hybrid CNN-BiLSTM model. If anything the paper reflected heavily on the significance of adopting hyperparameter tuning as observed by the outcome of both training and testing phase. The hyperparameter tuned CNN BiLSTM model offers a robust and effective approach compared to the traditional CNN models and underscores the importance of neural network architecture in capturing diverse features from the data.

The models could be further tuned with using computational heavy optimizers like Adam with Weight Decay, Rectified Adam or combining Adam and SGD similar to the approach taken in this paper. These optimizers although computationally heavy performs and yield better results in searching through the hyperparameters. The reason for not adopting this approach and going with a combination of Adam and Bayesian Optimizer is because these computational heavy optimizers often lead to faster convergence but better generalization for complex models like CNN especially when used in emotion analysis. The limitation of the hyperparameter tuning approach in this paper also suffers with faster convergence but it is still manageable both in results and computational resources.

References

Ramadhani, A.M.; Goo, H.S. Twitter sentiment analysis using deep learning methods. In Proceedings of the 2017 IEEE 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 1–2 August 2017; pp. 1–4.

Mika Viking Mäntylä, Daniel Graziotin, Miikka Kuutila The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers

Dholpuria, T.; Rana, Y.; Agrawal, C. A sentiment analysis approach through deep learning for a movie review. In Proceedings of the 2018 IEEE 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 24–26 November 2018; pp. 173–181.

Goularas, D.; Kamis, S. Evaluation of deep learning techniques in sentiment analysis from Twitter data. In Proceedings of the 2019 IEEE International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Istanbul, Turkey, 26–28 August 2019; pp. 12–17.

Pansy Nandwani. Rupali Verma A review on sentiment analysis and emotion detection from text

Jang, B.; Kim, M.; Harerimana, G.; Kang, S.U.; Kim, J.W. Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. Appl. Sci. 2020, 10, 5841.

Yang, Y. Convolutional neural networks with recurrent neural filters. arXiv 2018, arXiv:1808.09315.

Tyagi, V.; Kumar, A.; Das, S. Sentiment Analysis on Twitter Data Using Deep Learning approach. In Proceedings of the 2020 IEEE 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 18–19 December 2020; pp. 187–190.

Chundi, R.; Hulipalled, V.R.; Simha, J. SAEKCS: Sentiment analysis for English–Kannada code switchtext using deep learning techniques. In Proceedings of the 2020 IEEE International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 10–11 July 2020; pp. 327–331.

Rhanoui, M.; Mikram, M.; Yousfi, S.; Barzali, S. A CNN-BiLSTM model for document-level sentiment analysis. Mach. Learn. Knowl. Extr. 2019, 1, 832–847.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012 ImageNet Classification with Deep Convolutional Neural Networks.

Schusters, M., & Paliwal, K. K. 1997 . Bidirectional Recurrent Neural Networks.