Text Emotion Analysis

Anurag singh
CCE,LNM institute of information technology
Jaipur,India
21ucc020@lnmiit.ac.in

Mohit jain
CSE,LNM institute of information technology
Jaipur,India
21ucs131@lnmiit.ac.in

Himanshu Gautam
CSE,LNM institute of information technology
Jaipur,India
21ucs240@lnmiit.ac.in

ABSTRACT:
Emotional recognition has emerged as a crucial area of research, offering valuable insights across various domains. Emotions manifest in diverse forms, including speech, facial expressions, written text, and gestures. Detecting emotions in textual documents presents a content-based classification challenge, drawing upon principles from Natural Language Processing (NLP) and deep learning. This study proposes the use of deep learning-assisted semantic text analysis (DLSTA) for human emotion detection leveraging big data. Emotion detection from textual sources leverages NLP concepts, with word embeddings playing a pivotal role. Word embeddings are widely employed in NLP tasks such as machine translation, sentiment analysis, and question answering, enhancing learning-based methods by incorporating semantic and syntactic features of the text. To verify the performance of the proposed model, it is compared with previous studies. The optimal value is brought into the model contrast experiment. In the experiment, the accuracy rate, recall rate and F value of this method are all greater than 91%, which is the highest among the models. It effectively improves the accuracy of text sentiment analysis, and has high research and practical value.

1. INTRODUCTION

Text Classification (TC) refers to the systematic categorization of textual data into distinct groups based on their inherent properties and characteristics. Through automated analysis, TC effectively examined text and reliably assigned pre-established categories. This procedure is of utmost importance for facilitating the processing and extraction of valuable information from raw and unstructured textual data [1, 2]. TC encompasses three distinct system types: rule-based, machine-learning based, and hybrid approaches [3, 4]. Rule-based systems employ a collection of predefined rules to effectively categorize text into organized groups. On the other hand, machine learning-based systems rely on prior observations and patterns to classify text. Hybrid systems, as the name implies, combine elements of rule-based and machine learning-based systems, utilizing both trained classifiers to improve classification accuracy and efficacy.

Many researchers have dedicated their efforts to enhancing the efficacy of Natural Language Processing (NLP) tasks through the development of Deep Learning (DL) frameworks [5, 6]. Notably, CNN, Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM) techniques attained an accuracy rate of 77.4%, with CNN exhibiting the most favorable average performance [7]. These advancements in DL frameworks have contributed significantly to the improved performance and accuracy of NLP tasks in various applications. Researchers continue to develop hybrid DL models to achieve better results.

Combining CNN and BiLSTM in one architecture enables more comprehensive and in-depth data processing. CNN helps extract spatial features from the data, while BiLSTM helps extract temporal or contextual features [4]. Researchers have implemented hybrid CNN-BiLSTM models [9, 10], and using Word2Vec, GloVe, and fasttext. They used an embedding size of 300, Softmax activation, dropout, and recurrent dropout rates of 0.5 and 0.4, an SGD optimizer, 50 epochs, a filter size of 512, a kernel size of 3, embedding, CNN max pooling, BiLSTM, and dense layers, resulting in an accuracy of 82% [11].

Despite these advancements, challenges remain, including handling long inputs, unnecessary convolution operations for NLP, and data loss with increasing data size. The subsequent section of this article elaborates on the proposed methods, including data preprocessing, feature extraction, and model architecture. Experimental outcomes and dataset specifics are discussed in detail, followed by conclusions and insights into future research directions.

2. LITERATURE REVIEW

Previous studies related to text classification and sentiment analysis have been undertaken, and they adhere to the process of classifying text sentiment. These studies include:

1.The study aimed to enhance text classification accuracy using a Bi-LSTM model that combines Word2Vec, CNN, and attention mechanisms [11]. The approach combined multiple effective text classification techniques: Word2Vec for richer word vectors, CNN for local feature extraction, and attention mechanisms for weight assignment to significant sections of the text. The model used parameters such as a skip gram embedding size of 300, a 0.2 dropout rate, batch size of 128, and the Adam optimizer. It achieved an average accuracy of 87.4%, with the highest F1-Score (90.1%) observed on a 13k-instance dataset. This study demonstrated the synergistic integration of these models enhances text classification accuracy. Note that having access to more extensive training data and further refinement time benefits the model.

2.In a recent study, Salur and Aydin [9] introduced an innovative hybrid deep-learning model for sentiment classification. This model addresses dataset complexity and imbalances by synergistically combining CNN for local textual feature extraction and LSTM-based RNN for long term contextual information. To enhance feature representations, the model includes character embedding and pre-trained embedding (Word2Vec, GloVe, FastText). Using tweets related to a Turkish GSM operator, the results showed that a single BiLSTM model achieved 80.44% accuracy with FastText embedding. However, the hybrid CNN-BiLSTM model, which incorporates character embedding and FastText, outperformed it with 82.14% accuracy. These results emphasize the superiority of the proposed hybrid model for sentiment classification, although it is specifically tailored to Turkish, Arabic, and Lithuanian languages.

3.In a study by Naqvi et al. [10], sentiment analysis in Urdu text was addressed using deep learning. Their approach utilizes deep learning to enhance Urdu sentiment analysis accuracy, combining CNN for local feature extraction and RNN with LSTM for long-term context understanding. Four embeddings, Samar, CoNLL, pretrained, and self trained Fast Text, were evaluated. The BiLSTM ATT model achieved the highest accuracy at 77.9%, while LSTM reached the highest precision (85.16%) with Samar embedding. These results highlight the effectiveness of their deep learning methods in Urdu sentiment analysis. Further advancement potential exists through exploring alternative methods. This section reinforces the background by providing evidence and discussing trends in the field. It should encompass all relevant studies and supporting evidence.

A hybrid CNN-BiLSTM model was developed using different feature extraction techniques and kernels for the WAG dataset. This model is an innovative approach because it combines the advantages of both the techniques. The CNN is effective in its ability to extract local features from data, whereas the BiLSTM model excels in handling long-term dependencies and leveraging both past and future information to improve prediction accuracy. To accommodate varying data lengths, the inclusion of padding techniques adds value to the model's capability. The incorporation of the Rectified Linear Unit (ReLU) and Softmax activation functions further enhanced the flexibility of data processing. With an embedding size of 300, the model could extract more intricate and complex features from the input data.Following are the various state of the art models with their accuracies.

MODEL	**EMBEDDING**	**F1-SCORE**	**ACCURACY**
CNN LSTM CNN BiLSTM[7]	POS Tagging, sentiment vector	81.7%	77.4%
CNN BiLSTM [11]	Word2vec+ Skip Gram	88.0%	87.6%
CNN BiLSTM [9]	Carakter +Fastext	89.0%	82.1%

CNN BiLSTM CNN LSTM[13]	Sentiment Tagging	75.0%	85.5%

3. MATERIALS AND METHODS

This study aims to enhance the performance of the hybrid CNN BiLSTM model, specifically for sentiment analysis in the TC domain. We evaluated this hybrid model thoroughly by considering the advantages and potentials of both the CNN and BiLSTM architectures. The methodology we adopted for this study focuses on crucial aspects, such as selecting appropriate data, performing meticulous labeling, developing robust feature vectors, and formulating the hybrid CNN BiLSTM approach. All these aspects contribute to a more precise and accurate sentiment analysis solution.

A. Dataset

Dataset-1:
Dataset was combined from daily dialog, isear, and emotion-stimulus to create a balanced dataset with 5 labels: joy, sad, anger, fear, and neutral. The texts mainly consist of short messages and dialog utterances.
Test-https://github.com/Anu-vibes/text-emotion-analysis/blob/main/data_test.csv
Train-https://github.com/Anu-vibes/text-emotion-analysis/blob/main/data_train.csv

Dataset-2:
Emotions are expressed in nuanced ways, which varies by collective or individual experiences, knowledge, and beliefs. Therefore, to understand emotion, as conveyed through text, a robust mechanism capable of capturing and modeling different linguistic nuances and phenomena is needed. We propose a semi-supervised, graph-based algorithm to produce rich structural descriptors which serve as the building blocks for constructing contextualized affect representations from text. The pattern-based representations are further enriched with word embeddings and evaluated through several emotion recognition tasks. Our experimental results demonstrate that the proposed method outperforms state-of-the-art techniques on emotion recognition tasks.

https://huggingface.co/datasets/dair-ai/emotion/viewer

B. Labeling

Categorizing Labels: The original emotion labels ('joy', 'fear', 'anger', 'sadness', 'neutral') are mapped to integer labels using a predefined encoding dictionary (encoding). Each emotion label is associated with a unique integer value. This categorization simplifies the representation of emotion labels, making them suitable for feeding into machine learning models.

Integer Labels: The emotion labels in the training and testing datasets (data_train.Emotion and data_test.Emotion) are replaced with their corresponding integer labels using list comprehension. This transformation converts the emotion labels from their original string format to integer format, facilitating numerical computation and model training.

One-Hot Encoding: After converting the emotion labels to integer format, they are further encoded using one-hot encoding. One-hot encoding is a binary representation where each integer label is represented as a binary vector with a single '1' value at the index corresponding to the label's integer value and '0' values elsewhere. This encoding scheme ensures that the model interprets the emotion labels as categorical variables rather than ordinal variables, preventing unintended ordinal relationships between the labels.

By categorizing and encoding the emotion labels, we prepare them for consumption by the neural network model during training. This encoding scheme enables the model to effectively learn the mapping between input text data and corresponding emotion categories, facilitating accurate emotion classification.

C. Preprocessing

Preprocessing plays a crucial role in data cleaning as it significantly impacts subsequent analytical approaches. This process holds significant importance across diverse domains and languages. . Human-transferred text data needs to be in a machine-readable format for further analysis.Organizing a dataset requires preprocessing steps such as tokenization, stop word removal, POS tagging, stemming, and lemmatization. These techniques ensure data uniformity and remove irrelevant information for sentiment and emotion analysis.

Tokenization breaks down text into tokens, enabling the analysis of individual words or phrases. Normalizing text, including correcting spelling errors and converting text into standard form, enhances data consistency. Removing unnecessary words like articles and prepositions minimizes computational load during analysis. POS tagging identifies parts of speech, aiding in the extraction of relevant information for sentiment and emotion analysis.

Stemming reduces words to their root form by truncating suffixes, while lemmatization analyzes morphology to transform tokens into base word lemmas. Both processes optimize computational efficiency and improve analysis accuracy. Researchers have explored the impact of various preprocessing techniques on machine learning models' performance, highlighting the importance of steps like removing numbers and lemmatization for accuracy improvement.

D. Feature Extraction

In the field of NLP, a significant hurdle is developing a model capable of understanding the hierarchical representation of sentences in text data. This challenge arises primarily in the context of classification tasks and the extraction of relevant features . Feature extraction involves capturing the characteristic features or attributes of a particular shape and then carrying out an analysis of the captured feature values. The process of feature extraction involves reducing the dimensionality of the output data to make it more manageable and enable efficient processing. By carefully selecting and/or combining variables into characteristics, the amount of data to be processed is effectively reduced. Despite this reduction, the resulting features accurately and comprehensively characterize the underlying data. The data is organized into more manageable clusters. In this study, the following feature extraction processes were performed: Encoding is used to turn nominal or categorical data into numerical data by dividing the data into training and testing sets, tokenization to break down sentences into smaller pieces of words or tokens, and padding to adjust the output dimensions to match the input dimensions.

E. Model Used

The development of the hybrid CNN-BiLSTM model commenced with the labeling phase, pre-processing, and feature extraction. We employed an embedding approach to represent text as low-dimensional numeric vectors with a padding size set at 300,which differs from previous approaches as given.

Model: "sequential"

Layers	**Output Shape**	**Parameter**
Embedding Layer	(None, 200, 100)	7530300
Convolution1D Layer	(None, 200, 64)	19264
Pooling Layer	(None, 100, 64)	0
BiLSTM Layer	(None, 256)	197632
Dense	(None, 6)	1542

And also one more model as follows:

Model: "sequential"

Layers	**Output Shape**	**Parameter**
Embedding Layer	(None, 200, 100)	7514600
Convolution1D Layer	(None, 198, 200)	60200
BiLSTM Layer	(None, 198, 128)	135680
Dropout	(None, 198,128)	0
BiLSTM Layer	(None, 128)	98816

Dense	(None, 50)	6450
Dense	(None, 50)	2550
Flatten	(None, 50)	0
Dense	(None, 100)	5100
Dense	(None, 6)	606

F. Result and Observation

MODEL	DATA SET	EMBEDDING	ACCU RACY
ML(SVM)	1	Tf-idf vectorizer	72.71%
ML(SVM)	2	Tf-idf vectorizer	87.11%
BiLSTM	1	word2vec(wikipedia file)	72.68%
BiLSTM	2	Glove (twitter)	92.26%
CNN-LSTM (1)	1	word2vec(wikipedia file)	73.36%
CNN-LSTM (1)	2	Glove (twitter)	93.46%
CNN-LSTM (2)	1	word2vec(wikipedia file)	73.16%
CNN-LSTM (2)	2	Glove (twitter)	94%
BERT	1	Wordpiece	82.67%
BERT	2	Wordpiece	97%

CNN+BIlSTM[2] dataset 2 :

Emotion	precision	recall	f1-score
anger	0.94	0.95	0.94
fear	0.94	0.86	0.90
joy	1.00	0.91	0.95
love	0.77	1.00	0.87
sadness	0.98	0.97	0.97
surprise	0.75	0.99	0.85
accuracy			0.94
macro avg	0.89	0.95	0.92
weighted avg	0.95	0.94	0.94

4. REFERENCES

[1] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A.Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," IEEE Access, vol. 8, no. Ml, pp. 42689–42707, 2020. doi: 10.1109/ACCESS.2020.2976744

[2] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, "A systematic review of text classification research based on deep learning models in Arabic language," Int. J. Electr. Comput. Eng., vol. 10, no. 6, pp. 6629–6643, 2020. doi: 10.11591/IJECE.V10I6.PP6629-6643

[3] W. Fang, H. Luo, S. Xu, P. E. D. Love, Z. Lu, and C. Ye, "Automated text classification of near-misses from safety reports: An improved deep learning approach," Adv. Eng. Informatics, vol. 44, no. March 2019, 101060, 2020. doi: 10.1016/j.aei.2020.101060

[4] Z. Liu, C. Lu, H. Huang, S. Lyu, and Z. Tao, “Hierarchical Multi- granularity attention- based hybrid neural network for text classification,” IEEE Access, vol. 8, pp. 149362–149371, 2020. doi: 10.1109/ACCESS.2020.3016727

[5] Q. Li et al., “A survey on text classification: From shallow to deep learning,” IEEE Trans. NEURAL NETWORKS Learn. Syst., vol. 31, no. 11, pp. 1–21, 2020.

[6] F. Zaman, M. Shardlow, S. Hassan, and N. Radi, “HTSS : A novel hybrid text summarisation and simplification architecture,” Inf. Process. Manag., vol. 57, no. 6, 102351, 2020. doi: 10.1016/j.ipm.2020.102351

[7] K. Pasupa, T. Seneewong, and N. Ayutthaya, “Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding , POS-tag , and sentic features,” Sustain. Cities Soc., vol. 50, no. 7, 101615, 2019. doi: 10.1016/j.scs.2019.101615

[8] S. Kumar, C. Akhilesh, K. Vijay, and B. Semwal, “A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data,” Vis. Comput., vol. 38, no. 12, pp. 4095–4109, 2021. doi: 10.1007/s00371-021-02283-3

[9] M. U. Salur and I. Aydin, “A novel hybrid deep learning model for sentiment classification,” IEEE Access, vol. 8, pp. 58080–58093,2020.doi: 10.1109/ACCESS.2020.2982538

[10] U. Naqvi, A. Majid, and S. A. L. I. Abbas, “UTSA : Urdu text sentiment analysis using deep learning methods,” IEEE Access, vol. 9, pp. 114085–114094, 2021. doi: 10.1109/ACCESS.2021.3104308

[11] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, “Applied sciences Bi-LSTM Model to increase accuracy in text classification:Combining Word2vec CNN and attention mechanism,” Appl. Sci.,vol. 10, no. 17, 5841, 2020.

[12] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, “A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis,” Multimed. Tools Appl., vol. 78, no. 18, pp. 26597–26613,2019. doi: 10.1007/s11042-019-07788-7

[13] S. Soumya and K. V Pramod, “Hybrid deep learning approach for sentiment classification of malayalam tweets,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 891–899, 202