

***School of Computer Science Engineering and Information Systems***

**Department of Computer Applications**

**Fall Semester 2023-2024**

**SET CONFERENCE**

**First Review (11th October 2023)**

**Reg. No. -** 23MCA0304, 23MCA0310, 23MCA0360

**Student Name -** Sayantan Bhattacharyya, Milind Chakraborty, Rohan Raj

**Guide Name** – Prof. Nagalakshmi Vallabhaneni

**Project Title -** Comparative Study of Sentiment Analysis on Text data using various Deep Learning Models

**Problem Description:**

Social media applications have completely transformed the way people access information online. These platforms enable individuals to share their personal opinions on various topics like social events, product reviews, and film critiques. As a result, they generate a wealth of emotionally charged content. Gaining insights into this data can be useful in capturing public opinion and user interests. However, with the vast amount of information available, manual processing alone cannot accomplish this task. This led to the development of text sentiment analysis technology, which has become increasingly sophisticated over time. While traditional machine learning models like Naïve Bayes and SVM were once commonplace, deep learning models have since gained popularity due to their superior performance in various fields. There are several deep learning models available for sentiment analysis, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and hybrid models that combine different techniques. Each model has unique strengths and weaknesses, depending on the data and the task. Thus, a comparative study of these models is necessary. Our paper aims to explore the accuracy of different deep learning models in determining social sentiments using the IMDB dataset, which is widely recognized in the field.

**Literature Survey:**

Muhammet Sinan BASARSLAN and Fatih KAYAALP [1] present a sentiment analysis study which was carried out with 7 models based on various methods of deep learning algorithms on the IMDB dataset. The best result was obtained with the model consisting of 2 Bi-LSTM and 2 dropout layers with 80%-20% train-test separation and an accuracy value of 88.21%. They also observed that GRU or LSTM layers improved model performance, despite more Dense layers. Adam worked well with Activation. Bi-GRU alone gave the best results, despite Dropout's positive effect on performance.

Lenz Baron S. Balita et al[2]aim to construct a CNN-LSTM hybrid sentiment analysis model to examine social sentiments within Goodreads book reviews. The paper also aimed to determine the viability and effects of amalgamating features such as Word2Vec, POS, and SenticNet to the overall accuracy. Once the model was trained to the procured dataset, the results suggested that combining Word Embedding, POS, and SenticNet features drastically improves its performance in contrast to other tested variations. In terms of accuracy, the proposed hybrid CNN-LSTM model with Word2 + POS + SenticNet achieved 89.53% accuracy which outperformed single CNN with Word2Vec + POS(81.07%), single LSTM with Word2Vec + POS(84.27%), and CNN-LSTM hybrid with Word2Vec(86.07%) on the filtered GoodReads Dataset for book reviews.

Mahesh Mishra and Amol Patil [3] also constructed a hybrid model of CNN-LSTM on the IMDB dataset, where the LSTM architecture outperformed CNN and CNN-LSTM architectures. The accuracy rates of the GRU, CNN, LSTM, and CNN-LSTM architectures were 53%, 85%, 87%, and 85%, respectively.

Ling Zhang et al[4] study deep learning-based sentiment analysis methods, combining long and short-term memory network (LSTM) with an adaptive boosting algorithm (Adaboost), and propose a text sentiment analysis method based on LSTM-Adaboost. The IMDB movie review dataset is selected as the experimental object The results show that the LSTM-Adaboost model(87.47% accuracy, 89.51% precision, 84.9% recall and 87.14%) has the best result among the LSTM, CNN, and LSTM-Adaboost.

Rui Man et al[5] propose a sentiment analysis algorithm that combines BERT and CNN. It uses BERT to extract the features of each word and uses it as the input of CNN. After convolution and pooling, it is connected to the Softmax layer for classification. Here they used SVM, CNN, and BERT-CNN models on The hotel review corpus (ChnSentiCorp) and got the best accuracy in BERT-CNN at 90.5% which is far better than the other two models(81.3%, 85.2%).

K. Amulya et al[6] make the comparison of ML and DL approaches is done by considering IMDB movie reviews. From the observations, it is found that DL approaches provided more accurate results than ML algorithms. Among the DL algorithms (CNN, RNN, LSTM), RNN gives more accuracy of 88%.

Kavita Arora et al[7] used word2vec for word embedding and BERT(Bidirectional Encoder Transformers) model to classify and got an overall 92.40% accuracy to depict that the proposed scheme is an effective and reliable technique to detect sentiments for movie reviews.

Dr M Anusha and R Leelavathi[8] made a comparative study comparing deep learning models CNN-LSTM, CNN, and CNN-RNN on datasets including Sick SST, Twitter, FNC, Sentence Polarity, and IMDb. CNN-LSTM achieved an accuracy of 86.60% (Sick SST), 96.80% (Twitter), 97.80% (FNC), 98.60% (Sentence Polarity), 90.26% (IMDb), and 82% (English Data). Similarly, CNN achieved accuracy of 93% (sent-strength), 99.07% (sentence polarity), 94.80% (hostel data), 97.70% (IMDb), and 87% (SemEval), while CNN-RNN achieved accuracy of 89.67% (IMDb sentiment 140), 95% (KBP37), 86.60% (SST/DM/CS), and 94.60% (IMDb). The study addressed detecting sentiment polarity, pooling, ranking, and extraction.

Lirong Yao and Yazhuo Guan[9] propose an improved NLP method based on long short-term memory (LSTM) structure, whose parameters are randomly discarded when they are passed backwards in the recursive projection layer. Street Baseline achieved 86.51%, Classic LSTM 88.6%, Classic BiLSTM 89.3%, and improved LSTM got the best 90.17% accuracy on Wall Street Journal Dataset.

K. Mouthami et al. [10] utilized BiLSTM and GRU models on the IMDB dataset, finding that BiLSTM achieved the highest accuracy score of 99.65% after 10 epochs, while GRU resulted in an accuracy of 98.24%.

**Proposed Work:**

The study’s proposed conceptual framework begins with the procurement of data from the Amazon Review Polarity dataset to formulate the model’s training and testing dataset. Data cleansing comes right after data gathering wherein the pre-processing stage commences to segment the textual data, tokenize, remove unnecessary stop words, and stemming and/or lemmatization. Once the raw data are cleaned, the researchers segregate them as either a part of training or testing datasets. Once the cleansing stage is complete we create a model to extract text from reviews of movies and classify them as positive or negative according to the polarity of the reviews. We would compare various DL algorithms and consider the accuracy of each to make a comparison.

**Detailed Design:** **Architecture Diagram**

IMDB DATASET

Data Extraction

Data Preprocessing

Data Splitting

Train Test Split

Validation of the data by using various DL models

Hyper Parameter Tuning

Output Model Accuracy

**Methodology:**

**DATASET COLLECTION:**

We obtained our dataset from open-source resources, including Kaggle, to ensure a sizeable and representative corpus. The IMDB film review dataset met our requirements.

**DATA PREPROCESSING:**

1. **Text preparation:** Removing HTML tags, removing non-alphabet, making it lowercase
2. **Removing STOP words:** Eliminating common, non-informative words to reduce noise.
3. **Tokenization:** Splitting the text into individual words or tokens for analysis**.**
4. **Encoding and Embedding:** Transforming text into numeric form using various encoder or word embedding methods like word2vec.

**DATA SPLITTING:**

Train data is divided into two sets: the train set, used to train models with various weights, and the test set, used to evaluate the models' performance on untrained data.

**MODEL CREATION:**

Various models like LSTM and CNN are created, with additional layers and parameter values selected.

**MODEL TRAINING AND TESTING:**

After creating the models, we will train them with the train set, and test their performance using the testing set.

**HYPERPARAMETER TUNING:**

To further improve model performance, we can adjust hyperparameters such as the train-test split, optimizer, activation function, layer configuration, and add dropout layers etc.

**References:**

[1] Basarslan, Muhammet Sinan, and Fatih Kayaalp. "Sentiment Analysis with Various Deep Learning Models on Movie Reviews." In *2022 International Conference on Artificial Intelligence of Things (ICAIoT)*, pp. 1-5. IEEE, 2022.

[2] Balita, Lenz Baron S., Kyle Matthew A. Degrano, Andrei Daniel A. Pamoso, and Joel C. De Goma. "Sentiment Analysis on Book Reviews Using Convolutional Neural Network (CNN) Long Short-Term Memory (LSTM) Hybrid." In *Proceedings of the 2022 6th International Conference on E-Business and Internet*, pp. 125-132. 2022.

[3] Mishra, Mahesh, and Amol Patil. "Sentiment Prediction of IMDb Movie Reviews Using CNN-LSTM Approach." In *2023 International Conference on Control, Communication and Computing (ICCC)*, pp. 1-6. IEEE, 2023.

[4] Zhang, Ling, Miao Wang, Ming Liu, and Haozhan Li. "Sentiment Analysis of Movie Reviews Based on LSTM-Adaboost." In *2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, vol. 5, pp. 186-190. IEEE, 2022.

[5] Man, Rui, and Ke Lin. "Sentiment analysis algorithm based on bert and convolutional neural network." In *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pp. 769-772. IEEE, 2021.

[6] Amulya, K., S. B. Swathi, P. Kamakshi, and Y. Bhavani. "Sentiment analysis on IMDB movie reviews using machine learning and deep learning algorithms." In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 814-819. IEEE, 2022.

[7] Arora, Kavita, Neha Gupta, and Sonal Pathak. "Sentimental Analysis on IMDb Movies Review using BERT." In *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 866-871. IEEE, 2023.

[8] Anusha, M., and R. Leelavathi. "Analysis on sentiment analytics using deep learning techniques." In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 542-547. IEEE, 2021.

[9] Yao, Lirong, and Yazhuo Guan. "An improved LSTM structure for natural language processing." In *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*, pp. 565-569. IEEE, 2018.

[10] Mouthami, K., N. Yuvaraj, K. K. Thilaheswaran, and K. J. Lokeshvar. "Text Sentiment Analysis of Film Reviews Using Bi-LSTM and GRU." In *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1379-1386. IEEE, 2023.

1.

2.

3.

Signature (Students) Signature (Guide)

(Prof. Nagalakshmi Vallabhaneni)