IMDB Sentimental Analysis using Convolutional Neural Network

*

Individual Paper

Mashiya Farjana Jeba

19101394

Computer Science And Engineering

BRAC University

Abstract : Social networking is becoming a great way to get feedback from customers . With mood analysis , user reviews of a topic can be looked at and conclusions can be drawn.

based on what the reviews showed about how people felt about the product.Deep neural networks like the convolution neural network (CNN) and the recurrent neural network (RNN) are often used to analyse how people feel.We looked at CNN and RNN models for rating the emotional content of IMDb movie reviews in order to figure out which one would work best for the dataset .The CNN model that was learned with these datasets was correct 99.18% of the time during training and 85.06% of the time during testing. But the RNN model that was made using these datasets was accurate 82.84% of the time.

Index Terms : sentiment analysis , IMDb movie reviews , CNN , RNN, and NLP.

I. Introduction

Sentiment analysis is one of the most popular and widely used research tools , and new areas of study include text mining and natural language processing (NLP). It is quickly becoming one of the most important and interesting areas of study because a product's success depends a lot on how well it is reviewed online. Analysis of mood helps us understand the relationship between how people feel and what they say or write about natural surroundings or judgement.

It is beneficial for us to examine a person's viewpoint regarding an organization that holds a significant lot of significance for the creator of the entity. For illustration , in our day and age, almost no one goes to the movies .That is, unless they had seen some positive reviews of the movie on social media. Media sources or a few film critics No matter if you are purchasing goods or services, the terms are the same. As a result, reviews are starting to displace the world of promotion. Decrease the amount of error and difficulty in sentiment prediction as a result , as this is something that should be substantial and vital.Behind the scenes of a review .Differences from earlier eras techniques in machine learning, such as Nave Bayes and support vector machines [1], employed in a sentiment analysis with satisfactory outcomes . Due to its compositional capabili ties and local in variance,One of the most well- known neural network architectures for image classification is CNN .. Additionally , CNN has demonstrated natural language processing , which is significant for its distinctive identification method since it can quickly identify the key term needed for sentiment analysis in the natural text.. We must determine if people will experience positive or terrible feelings in order to This is something CNN undertakes to discover what tastes people have. for us. Three levels CNN has been used by Ouyang, Zhou, and Li [2] on the dataset of movie reviews from the rotten tomatoes website. For the classification of texts, Kim [3] has also used a variety of Convolution Neural Network models. CNN is smaller Training takes less time as a result of this relationship, which is a great benefit.

II. Literature Review

The study of opinion, sentiment, and subjectivity in text using machine analysis is known as sentiment analysis [4].So cial networking platforms like Facebook, Twitter, etc. contain a vast amount of data. Due to this, SA is a difficult task, and analyzing social media content raises several problems[5].In [6], CNN and Word 2Vc were combined .An outline has been suggested . In order to increase the generalizability and ac- curacy of the Word 2vec model used to read the text in the movie reviews dataset, the authors proposed a seven-layer model.. The accuracy of the proposed model, which combines the CNN model, PReLU, and dropout technology , is 45.4%, which is higher than that of another neural networkThe relaxed online maximum margin approach [7] selects a hyper-plane repeatedly that accurately categorizes the training examples that are now available with the maximum margin . The trade - off between the amount of progress gained on each training round and the knowledge maintained from earlier rounds is managed by the Passive Aggressive (PA) algorithm [8].The study [13] collected opinions and perspectives on major events from online users by using CNN comments on microblogs. Because the CNN method defeats feature extraction and im- plicitly learns the data, it was used.In total 1000 comments from microblogs were collected and categorized using three distinct categories. The Deep Belief Network with Feature Selection is the au thors' remedy for the vocabulary issues in [10]. (DBNFS).The given approach effectively resolves the polarity problem for a dataset . Additionally, authors proposed a weighted K closest neighbor classifier (Weighted K-NN) that performed better than the K nearest neighbors in terms of accuracy . Weighted K NN classifier evaluations from online review sites like flipcart, ebay, and amazon.com that are strongly polarized correctly classify weekly and light data. An illustration . The proposed classifier allows the flexibility to adjust the parameters to meet system needs. In [11],it is advised to use a Gini Index features selection technique with SVM to categorize the viewpoints stated in movie reviews. Turney [12] discovered sentiments in texts that followed pre- established part-of-speech patterns.

III. Research Motivation and Objective

1) Motivation: The motivation for using RNNs and CNNs for sentiment analysis on IMDb data could be to gain insights into the public perception of movies and to improve the user experience on the IMDb website. By accurately predicting the sentiment of movie reviews , the website could recommend movies to users based on their preferences and provide studios with valuable feedback on the public's reaction to their films. Additionally, sentiment analysis can help identify patterns and trends in movie reviews, which could be useful for conducting market research and improving the movie industry as a whole.

2) Objective: The objectives of using RNNs and CNNs for sentiment analysis on IMDb data would be to accurately predict the sentiment (positive or negative) of movie reviews. This could be useful for various applications, such as improving the user experience on the IMDb website by recommending movies based on the sentiment of reviews, or providing feedback to movie studios on the public perception of their films.

IV. Proposed Methodology

1) Input data: The dataset was obtained from Kaggle[13]. There are 50,000 movie reviews in the dataset, divided into two groups: positive and negative. This dataset, which displays each review as a collection of word indexes, was provided to us by kaggle[13].

2) Pre-Processing: Preprocessing data is essential for the CNN approach. This approach eliminates variables that don't improve CNN's accuracy or results.This stage enables us alter the raw data to improve the CNN model's performance and accuracy. To construct our method for detecting reviews, we separated our image collection into folders like: positive and negative. Our training dataset was also updated. There were many variables employed, including word embedding, batch size, maximum length, and word count. Each phrase has a distinct ID, and the pre-processed embedding layer dataset provides a distinctive and significant word order. After initializing the words with random weights,Each word in the training dataset is embedded by the embedding layer through learning.. Although this layer has many various applications, its primary use is to identify word embeddings that can be retained and used in other models.

Convolutional Neural Network(CNN):-

- Word Embedding : 32
- Batch size : 128
- Max length : 160
- Number of Words : 25000

Recurren Neural Network(RNN):-

- Word Embedding : 32
- Batch size : 128
- Max length : 160
- Number of Words : 25000

A. Proposed CNN Architecture

1) Implementation: We built a CNN model that can recognize positve and negative reviews from our datasets. This classification issue will use data attributes to uncover distinct patterns and distinguish between datas, identifying revies. Graphic processing power is needed to classify revies. So, a GPU is needed. We needed a discrete GPU because it handles graphic processing. Python is the most popular programming language for this assignment, so we're utilizing it.

2) CNN Model Summary: The number of parameters in a layer represents the filter's "learnable" (if such a concept exists) components. We constructed a sequential CNN model for the specified system using the keras neural network toolbox. The model's accuracy was tested. Model has 9 layers.There were also included three 1D convolutional layers, and one max - pooling and one layer of batch normalization.In our model, we employ the RELU activation function. When the value is negative, Relu returns 0, and when it is positive, it increases. The dense layer, also known as the fully connected layer, is used to categorize the characteristics gathered by the convolutional layers. Using dense layer, every input (or neuron) in the layer of the network that is currently active is connected to every input (or neuron) in the layer that comes after it.Two layers of dense paint, then one layer of flatten. An activation function is really just a straightforward function that transforms its inputs into outputs with a specified range. For instance, the sigmoid activation function accepts input and generates output values that range from 0 to 1. We collected 818,937 trainable parameters for the model to use when training data.

TABLE I
TABLE OF CNN MODEL

Layer	Output Shape	Param#
embedding (Embedding)	None,160, 32	800000
conv1d (Conv1D)	None,158, 64	6208
conv1d_1 (Conv1D))	None,156, 32	6176
conv1d_2 (Conv1D)	None,154, 16	1552
max_pooling1d (MaxPooling1D)	None,77,16	0
batch_normalization (BatchNormalization)	None,77,16	64
flatten (Flatten)	None,1232	0
dense (Dense)	None,4	4932
$\text{dense}_1(Dense)$	None,1	5
Total params:		818,937
Trainable params:		818,905
Non-trainable params:		32

B. RNN Architecture

1) Implementation: We built a RNN model that can recognize positve and negative reviews from our datasets. This classification issue will use data attributes to uncover distinct patterns and distinguish between datas, identifying revies. Graphic processing power is needed to classify revies. So, a GPU is needed. We needed a discrete GPU because it handles graphic processing. Python is the most popular programming language for this assignment, so we're utilizing it.

2) RNN Model Summary: The number of parameters in a layer represents the filter's "learnable" (if such a concept exists) components. We constructed a sequential RNN model for the specified system using the keras neural network toolbox. The model's accuracy was tested. Model has 3 layers.There were also included one embedding layer, and one one simple rnn layer. one layer of dense paint. We collected 800,801 trainable parameters for the model to use when training data.

TABLE II
TABLE OF RNN MODEL

Layer	Output Shape	Param#
embedding (Embedding)	None,160, 32	800000
simple_rnn (SimpleRNN)	None 16	784
dense (Dense)	None,1	17
Total params:		800,801
Trainable params:		800,801
Non-trainable params:		0

V. RESULT & ANALYSIS

A. Proposed Model Analysis

We used a total of 50000 data for our bespoke model.The data were categorized into two groups: "Positive," as well as "Negative."The proposed model had a 85.06% accuracy rate.The table below shows the recall, precision, training accuracy, and validation accuracy of our suggested model.As mentioned earlier, our training accuracy was 99.18% and our testing accuracy was 85.06%. From the below figure 2 we can see how fast training accuracy increased with time

TABLE III
ACCURACY OF PROPOSED MODEL

Training accuracy	Validation accuracy	Precision	Recall
99.18%	85.06%	99.98%	98.38%

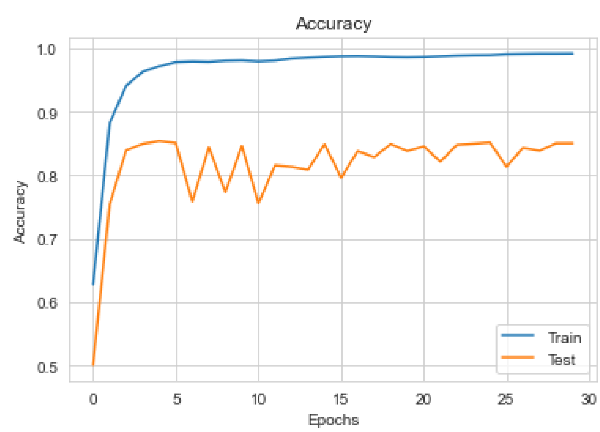

Fig. 1. The suggested model's training and validation efficiency graph

B. Comparison & analysis for all model

In this part, we compared the proposed CNN model to RNN.The proposed model beats the RNN model, as shown in Table 1. Table 8.1 compares RNN and proposed model

TABLE IV
COMPARISON BETWEEN ALL FRAMEWORKS

Name of the models	Training Precision	Validation Precision
Proposed Model	99.18%	85.06%
RNN	99.88%	82.84%

accuracy. Proposed model has the best accuracy, 85.06%. RNN has 82.77

As the training accuracy of RNN is low comparative to the proposed model. That's why it has the low accuracy.We can say that proposed model achieved better performance than RNN model. Equal amount of datasets were utilized to determine the model accuracy. The pie chart given below shows the accuracy of the models.

The pie chart given below shows the accuracy of the models.

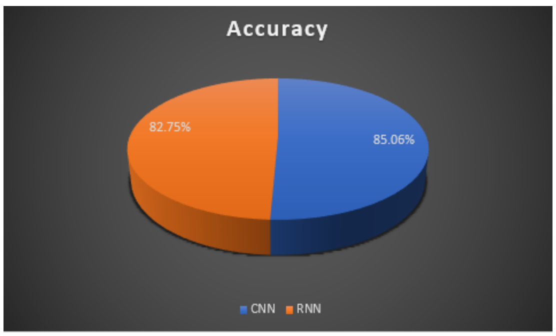

Fig. 2. Comparison of all models efficiency

C. Efficiency Evaluation on Models

We generate the confusion matrix for both CNN and RNN model. The CNN model successfully predict 10167 true positive,11098 true negative,2333 false Positive and 1402 false Negative.

On the other hand,the RNN model successfully perdict 10360 true positive,10334 true negative,2140 false positive and 2166 false negative.

VI. CONCLUSION & FUTURE WORK

Learning how to extract characteristics from the data with the aid of CNN.Given how quickly the volume of online data is growing, sentiment analysis is becoming increasingly crucial.We want to detect negative and positive reviews us ing CNN and RNN. CNN model was also trained. Which detected negative and positive reviews with 85.06% accuracy. To improve our model's accuracy, we'll train it using diverse datasets. We aim to solve the problem. This model aims to predict reviews more efficiently

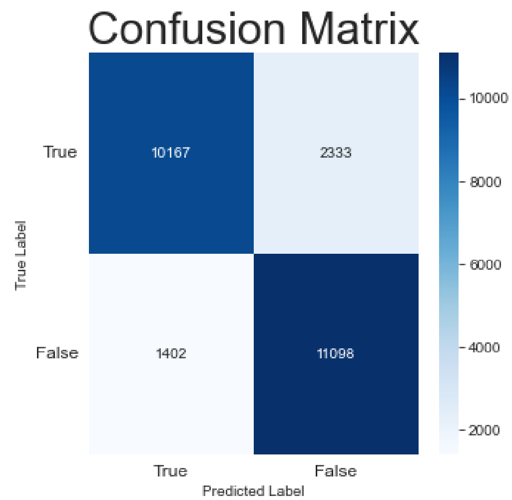

Fig. 3. Confusion matrix of CNN model

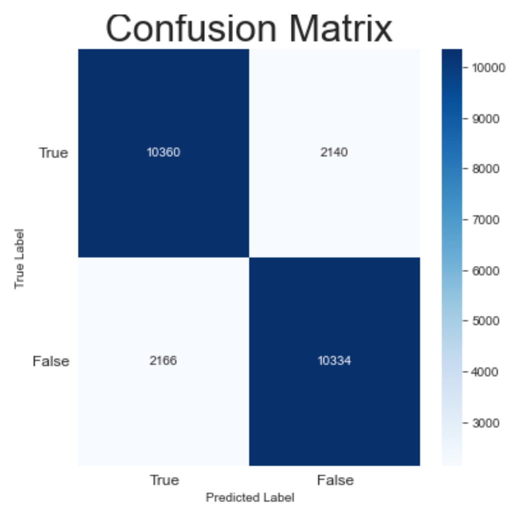

Fig. 4. Confusion matrix of RNN model

REFERENCES

[1] Rana, S., Singh, A., ""Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques." 2016 2nd International Conference on Next Generation Computing Technologies (NGCT). IEEE, 2016, pp. 106-111 ".

[2] Ouyang, X., Zhou, P., Li, C. H., Liu, L .,""Sentiment analysis using convolutional neural network." 2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing. IEEE, 2015, pp. 2359-2364.

[3] Kim, Y .,""Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882, 2014.

[4] Pang and L. Lee., "Opinion mining and sentiment analysis Found Trends Inf. Retr, 2(1-2):1–135, 2008,".

[5] Li G, Hoi SC, Chang K, Jain R (2010) .,"Micro-blogging sentiment detection by collaborative online learning. In IEEE 10th International Conference on Data Mining (ICDM), pp. 893-898.

[6] Pang B, Lee L (2005) .,") Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of the 43rd annual meeting on association for computational linguistics, pp. 115-124.

[7] Y. Li and P. M. Long.,"The relaxed online maximum margin algorithm. In NIPS, pages 498–504, 1999.

[8] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer.,passive-aggressive algorithms. J. Mach. Learn. Res., 7:551–585, 2006.

[9] Yanmei L, Yuda C (2015) .,"Research on Chinese Micro-Blog Sentiment Analysis Based on Deep Learning. In IEEE 8th International Symposium on Computational Intelligence and Design (ISCID), pp. 358-361.

[10] Ruangkanokmas P, Achalakul T, Akkarajitsakul K (2016) .,"Deep belief networks with feature selection for sentiment classification. In IEEE 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 9-14.

[11] Conneau A, Schwenk H, Barrault L, Lecun Y (2017) ,"Very deep convolutional networks for text classifi cation. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1107-1116.

[12]] P. D. Turney.,"TThumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In ACL, pages 417–424, 2002.

[13] https://www.kaggle.com/datasets/vikramtiwari/imdb-dataset-for-keras-imdbnpz?select=imdb.npz