Multi-class Sentiment analysis of Rotten Tomatoes movie review corpus using CNN

Preetkumar Patel
Lakehead University

I. ABSTRACT

Old text sentiment analysis techniques often neglect the context when it is used in the expression of features. In this research, we have implemented a scalable and robust CNN (Convolutional neural network) to address the issue of text-based film review sentiment analysis. The proposed method classifies the sentiment of sentences in the Rotten Tomatoes data set. Basically, it is a multi-class classification, which means that the data set have more than 2 classes for classifying the movies. Finally, we are evaluating the model on testing dataset to check how it performs by calculating Accuracy, Recall, Precision and figure of merit measures. Moreover, this paper also discuss about the previously implemented research work on sentiment analysis which uses different methods on various dataset.

Keywords–CNN, Sentiment Classification, LSTM, Relu, Sigmoid, Max pooling, SVM, precision, recall, F1 Score

II. INTRODUCTION

The natural language processing function involves sentiment analysis, paraphrasing, recognition of entailment, summarization, discourse analysis, grounded language learning, retrieval of images and machine translation. Sentiment Analysis is a Natural Language Processing (NLP) technology that is used on the Internet to find the emotions of user's feedback, comments etc. A new film is released every year; it'll be easier if there's a recommendation to screen the film that people want to watch. There are a number of suggestions for films today, one of them being rotten tomatoes movie review corpus. Mainly sentiment analysis can be done using two approaches, lexical approach and supervised machine learning methods. Convolutional neural networks (CNN) and long term memory networks (LSTM) are commonly used in the text classifications which are deep learning algorithm. CNN is a type of multi-layer neural network which is considered to be improvement of error back propagation network. The basic structure of CNN is shown in the Figure 1. Usually, bag-of-word is used for representing the text into fixed length vector and then applies CNN in order to classify text. Usually, bag-of-word consists of vocabulary of known words and a measure of presence of known words. It results in a form called a sparse form or sparse representation, with lots of zero scores. When modeling, sparse vectors need more memory and computing resources and the large number of positions or dimensions can make the modeling process very difficult for conventional algorithms. However, bag-of-words is not so useful because it does not care about the context of the text.

Hence in the proposed model, we have used TF-IDF Vectorization for feature extraction. Here, term frequency is a measure of how often a term appears in a document whereas inverse document frequency means gives importance of the word. The concept behind TF-IDF is to give greater priority to the terms that appear more frequently in one document and less frequently in other documents, because they are more useful for classification.

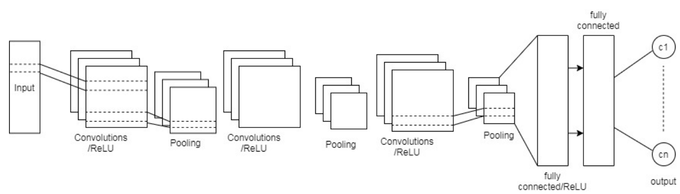

Fig. 1. CNN architecture with convolutional layers which are continuously stacked between ReLus before passing through the pooling layer, before going between one or more fully connected ReLus.

III. RELATED WORK

There are many techniques which are developed for multi-class sentiment classification. Techniques for sentiment analysis can be split into rule-based methods and statistical-based methods. At present, Machine learning is actually the primary approach used for evaluating sentiments. One such method was developed in [1] which combines CNN and LSTM (long short term memory recurrent neural network) for deep sentiment analysis. In this model, two CNN layers and one layer LSTM is used. This model gives accuracy of 78.42% which is better than accuracies of single CNN, SVM and CNN-LSTM model [1].

In recent years, several studies have concentrated on behaviors with the help of emojis. Many of these research used emojis to ease issues such as interpretation of thoughts, classification of subjectivity, classification of polarity, classification of emotions, and identification of ironies. In study [2], using two separate representations based on bag-of-word and fastText, they used and tested classification models to tackle the issue of sentiment analysis over emojis for positive, negative and neutral tweets. The label of each tweet is categorized as emojis [2]. This model is considered as emoji prediction model. BoW vectors are trained in four separate classificators: Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), and Decision Tree (DT). For comparison, fastText is used, which shows better results as compared to BoW.

IV. Dataset Discription

This is a movie review dataset which is a collection of film reviews and it collected by Pang and Lee in [3].Moreover, the dataset was analyzed in [4] where each sentence is represented in its tree structure and each node is assigned a sentiment label which range from 1 to 5 where the numbers are very weak, weak, neutral, positive and very positive respectively. It includes attributes such as PhraseID, SentenceId, Phrase and Sentiment as show in the Figure 1. The figure shows saved image of the tail of the dataset and it also indicates that the dataset is unstructured.

	PhraseId	SentenceId	Phrase	Sentiment
156055	156056	8544	Hearst 's	2
156056	156057	8544	forced avuncular chortles	1
156057	156058	8544	avuncular chortles	3
156058	156059	8544	avuncular	2
156059	156060	8544	chortles	2

Fig. 2. Unstructured dataset

V. Proposed Methodology

The aim of the proposed method is multiclass sentiment analysis using CNN on rotten tomatoes movie review data set. The basic steps are shown in the Figure 3. After reading the dataset, full sentences of the dataset are extracted and kept in Phrase column. After this, the result obtained is shown in the Figure 4. This is done because some of the values in phrase column are only words of the full sentences. So it is necessary to get full sentence before performing further task. Next step is to slip the dataset into train and test data where training dataset is taken as 70% and testing dataset is taken 30%. After splitting the dataset preprocessing steps are performed. Pre-processing is necessary because the dataset taken for sentiment analysis is unstructured and it is necessary to clean the text before performing any technique for sentiment classification. There are various steps performed to clean the dataset. In data preprocessing, punctuations are removed because it is useless in classification purpose. Secondly, all the words are converted to lower case and stop words are removed.Moreover, TF-IDF vectorization is performed for feature extraction.

After preprocessing step, the conv1D model is defined and trained on the training dataset. This is used for prediction of the given dataset. The follwing layers are used in the approach:input layer, max pooling layer, convolutional layer, flatten layer, linear layer, output layer. In this, output of the first layer is obtained and run through the ReLU activation function. After this,the output of the second layer is obtained and again run it through the ReLU activation function.Basically, three convolutional layers are used in this model. Now, get

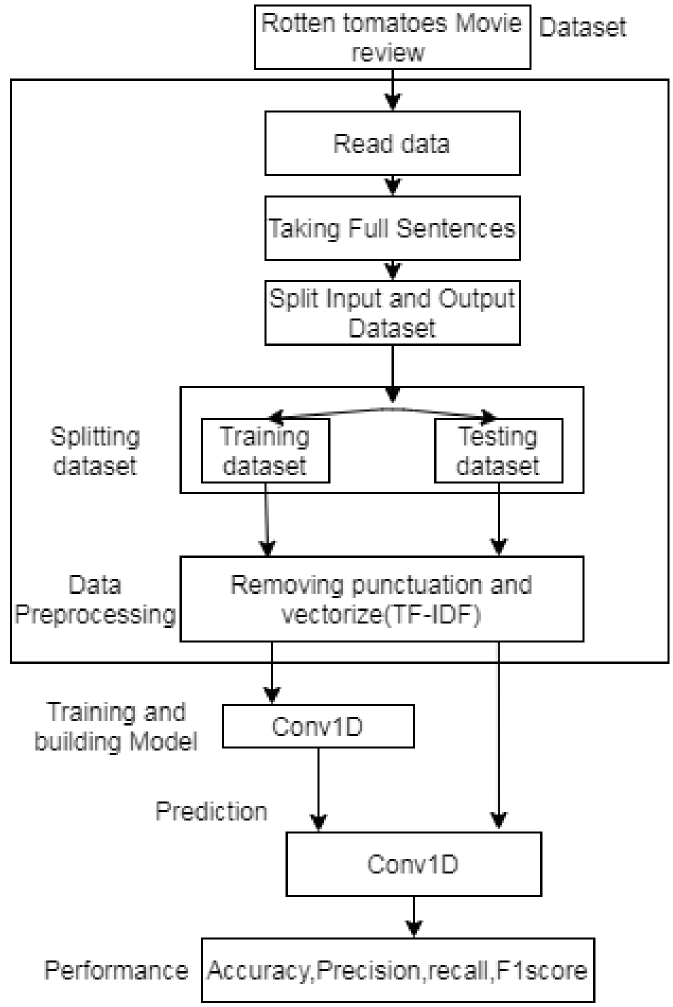

Fig. 3. Flow diagram of proposed method

	Phrase	Sentiment
0	A series of escapades demonstrating the adage that what is good for the goose is also good for the gander , some of which occasionally amuses but none of which amounts to much of a story .	1
1	This quiet , introspective and entertaining independent is worth seeking .	4
2	Even fans of Ismail Merchant 's work , I suspect , would have a hard time sitting through this one .	1
3	A positively thrilling combination of ethnography and all the intrigue , betrayal , deceit and murder of a Shakespearean tragedy or a juicy soap opera .	3
4	Aggressive self-glorification and a manipulative whitewash .	1

Fig. 4. Full sentences

the output of the flatten layer and linear layer and run it with activation function. Moreover, obtain the output of the output layer and return the output.Furthermore, model is trained and in this way, the steps to train the model are completed.After training, the performance of training model is evaluated on the basis of accuracy, precision, recall and F1 Score. Furthermore, after completion of the training, the models are used for obtaining the predictions on the test set. The model is then evaluated on the basis of the outcome from the prediction. This is the basic flow of the proposed method. The CNN model used consist of the following layers:

1) Convolution layer: There is a feature detector this layer. This is used to detect edges or specific shapes.

2) Pooling layer: This layer is used for parameter reduction and process computation. Invariant features are often identified by using this layer to scale or orientation shifts, so it avoids overfitting. There are two types of pooling,max pooling and average pooling. Mainly, max pooling is used in the many techniques. Both the types are shown in the Figure 5.

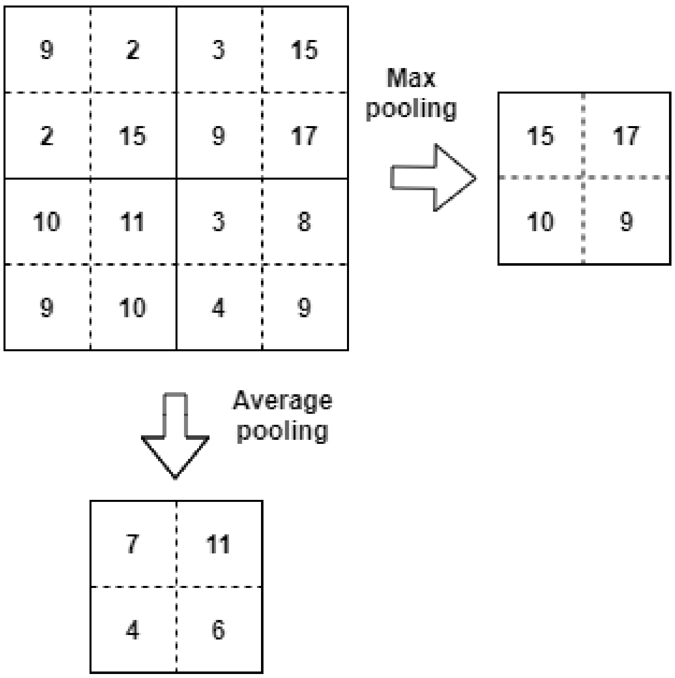

Fig. 5. Pooling

3) Flatten layer: Basically flattening is taking matrix from convolutional and pooling layer and transforming it into a one-dimensional collection.It is necessary step because input to fully connected layer need to be one dimension array.The flatten layer is shown in the Figure 6.

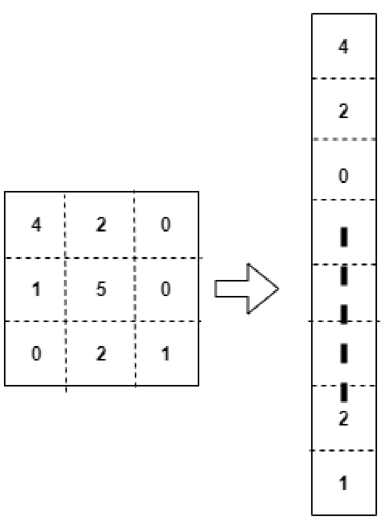

Fig. 6. Flatten Layer

4) Dropout: Dropout layer is used to reduce overfitting.It dropout hidden nodes in the network.
5) Dense Layer: A linear operation, in which each input is related by a weight to each output.

Other parameters which are used in the layers are:

- Activation Function: Activation function determines whether to activate a neuron or not by measuring weighted sum and adding bias with it further. The activation function has the role of integrating non-linearity into a neuron output.There are many activation functions. In the proposed model, relu in convolutional layer and softmax in dense layer are used.
- Optimizer:It updates the weights parameters for minimizing the loss function. In the proposed model, we have used two optimizers sgd(Stochastic gradient descent optimizer) and adam.
- Learning rate: The learning rate is a hyperparameter that every time the weights of the model are changed, it generally, determines how much to adjust the algorithm in response to the actually predicted error.Basically, Learning rate is between 0 to 1. For sgd optimizer, learning rate is 0.01 and for adam learning rate is 0.001.
- Kernel Size:A kernel is a part integral of neural network.Basically it is a Filter.In our model kernel size used is 3. In general for image, it refers to an operator applied to the entire image so that the information contained in the pixels is transformed.

VI. Results

The CNN model used in our methodology for the multi class sentiment analysis is:

Input(Embedding Layer)
model2.add(Embedding(max_feature,150,inptlen=maxwrds))
CNN Layers
model2.add(SpatialDropout1D(0.2)) Convolutional layer 1
model2.add(Conv1D(32, kernel_size=3, padding='same', activation='relu')) model2.add(MaxPooling1D(pool_size=2))
Convolutional layer 2
model2.add(Conv1D(32, kernel_size=3, padding='same', activation='relu')) model2.add(MaxPooling1D(pool_size=2))
Convolutional layer 3
model2.add(Conv1D(64, kernel_size=3, padding='same', activation='relu')) model2.add(MaxPooling1D(pool_size=2))
Flatten layer
model2.add(Flatten())
Dense Layer
model2.add(Dense(15, activation='softmax'))
Output layer
model2.add(Dense(5, activation='softmax'))

For the CNN model different parameters are used. The important parameters used in the model are shown in the Figure 7.

Moreover, optimizers used for the methodology are sgd and adam. Comparing both, it is seen that sgd is giving

Parameters	Values
No. of Epochs	25
Learning rate of sgd	0.01
Learning rate of adam	0.001
Kernal Size	3
Batch Size	64
Splitting of dataset	70:30
Pooling size	2

Fig. 7. Important Parameters

low accuracy whereas adam gives more accuracy. The adam optimizer performs better on the larger dataset as compared to sgd. However, in terms of computational speed sgd performs better than adam.The comparision between both of them is shown in the Figure 8. Furthermore, the performance is also evaluated on the basis of Precision, recall and F1 Score which is shown in Figure 9 and Figure 10.

Optimizer	Accuracy	No of Epochs
SGD	37%	25
ADAM	83%	25

Fig. 8. Accuracy Comparision

	Precision	Recall	F1 Score
0	0.88	0.98	0.93
1	0.83	0.77	0.83
2	0.77	0.69	0.73
3	0.78	0.76	0.77
4	0.88	0.96	0.92

Fig. 9. Performance matrix with adam

VII. Conclusion

The paper discuss about the multi class sentiment analysis of Rotten Tomatoes movie review dataset.The methodology used for classification is the Conv1D neural network. Moreover, the performance of the model is evaluated using precision, accuracy, recall and F1 Score as discussed in the results. All in all, it can be concluded that adam optimizer gives better accuracy than sgd optimizer when run on the given dataset.

	Precision	Recall	F1 Score
0	0.36	0.83	0.50
1	0.29	0.09	0.14
2	0.53	0.69	0.59
3	0.31	0.16	0.29
4	0.20	0.11	0.14

Fig. 10. Performance matrix with sgd

References

[1] B. Chen, Q. Huang, Y. Chen, L. Cheng and R. Chen, ”Deep Neural Networks for Multi-class Sentiment Classification,” 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, United Kingdom, 2018, pp. 854-859.

[2] R. Velioğlu, T. Yıldız and S. Yıldırım, ”Sentiment Analysis Using Learning Approaches Over Emojis for Turkish Tweets,” 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, 2018, pp. 303-307.

[3] Pang and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In ACL, pages 115–124.

[4] Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Chris Manning, Andrew Ng and Chris Potts. Conference on Empirical Methods in Natural Language Processing (EMNLP 2013).