Multi-class Sentiment Analysis using Deep Learning

Sonal Yadav
Computer Science
Lakehead University
Thunder Bay, ON
syadav3@lakeheadu.ca

***Abstract*—In this experiment a 1D convolution neural network (CNN) for text-based movie review is proposed. The movie dataset used is a multi class sentiment data by rotten tomatoes. The network architecture proposed is composed of layers like convolution, dropout layer, max pooling and fully connected layer. The advantage of the proposed model is that is predicts the multi class sentiment after preprocessing of the given data, with resampled data. The report further describes about reading the data, preprocessing the data by removing stop words and using lemmatizes and later using tokenizer to generate a input for the model to generate a desired accuracy, precision and F1 score.**

***Index Terms*—CNN, precision, F1 score, Word2Vec, Bag of Words**

I. INTRODUCTION

Sentiment analysis is formally defined as the task to identify and analyze subjective information of people's opinions in social media source. Many users have different opinions about products or movies and brands that impact other users before buying the product or watching a movie. These reviews have high impact, hence must be analyzed using proper techniques which in turn helps other users to decide fairly before making any decision. This field of study has attracted many researcher and developers to develop an optimized model which overcomes all the challenges and generate a true prediction. There can arise various challenges while building a optimized model which predicts the review truly. Some of the challenges are variety of expressions which can lead to various sentiments, which is difficult to handle even with a predefined dictionary. The words used to describe any product can have different meaning which can mislead and give a wrong decision. Not only words but also indirect reviews can sometimes generate a wrong decision about any product if not handled properly. To overcome such challenges sentiment analysis broadly classifies all the sentiments into 3 categories like positive, negative and neutral. To handle such multi class text sentimental data, we have proposed a 1D CNN model which fairly treats the data and generates a desired precision. Convolution Neural Network is basically used for 2D data classification such as images. A CNN is usually a feature Convolution Neural Network is basically used for 2D data classification such as images. A CNN is usually a feature map that corresponds to the features of the input. In text analysis the words in a sentence or a document or sequence of characters. CNN works well with text data also even though its one form of data and features can be extracted. The convolution layer is followed by a max pooling layer which helps in extracting maximum features and reducing the dimension of the input to that layer. The CNN network basically involves two main operations which are convulsion and pooling. The output of these operations are fed to the fully connected layer which works the same as any traditional neural network. Map that corresponds to the features of the input. In text analysis the words in a sentence or a document or sequence of characters. CNN works well with text data also even though its one form of data and features can be extracted. The convolution layer is followed by a max pooling layer which helps in extracting maximum features and reducing the dimension of the input to that layer. The CNN network basically involves two main operations which are convulsion and pooling. The output of these operations are fed to the fully connected layer which works the same as any traditional neural network.

In this report we focus on creating a model which predicts the multi class sentiment with the preprocessed data. Use of Bag of words is also done which further simplifies in preprocessing the[] data and focus only the words that truly define thee sentiment of the movie, The CNN model consists of various layers like max pooling and convolution which are then fed to full connected layer. The predicted results of the model are then used to calculate the required precision, accuracy and F1 score.

Identify applicable funding agency here. If none, delete this.

II. BACKGROUND

Sentiment Analysis is a major subject in machine learning which aims to extract subjective information from the textual reviews. The reviews predicted can tell us in general if a movie is positive or negative. It can also be used to determine the state of mind of the person whether the person was impressed with movie by positive feedback or was not happy with the movie. The field of sentiment of analysis is closely tied to natural language processing and text mining. Collecting data for sentiment analysis in a resource limited setting requires making the most out of all available data sources. These can vary wildly in terms of size, class distribution, and various linguistic properties. Thus, the collected data is likely to be imbalanced and to suffer from high sample selection bias, as it is not representative of the whole population[2]. The basic idea with deep learning is to use hidden layers of neural nets to automatically capture the underlying factors that lead from the input to the output, eliminating the need for feature engineering. Text mining is analysis of data containing the natural language. It refers to mining of the raw text data and deriving feature and information out of it. The data derived is basically information which is used for further classification. It helps the human to make optimized decisions. Text categorization on the other hand is classifying the document into one or more predefined categories. Such classification is possible with the help of deep learning algorithms which effectively classify into correct categories. The major hand behind optimized categorization is with the labels which help the classifier judge the category of the text. L. Terveen, designed a system uses a collaborative filtering approach to recognize and reuse recommendations. J. Tatemura, developed a browsing method using virtual reviewers for the collaborative exploration of movie reviews from various viewpoints [1]. Morinaga et al, presented a framework for mining product reputation on internet, the defined approach automatically collects the users opinions from the web and applies text mining techniques to obtain the reputation of the products[1]. Convolutional Neural Networks (CNNs) are a powerful deep learning technique because they preserve the spatial structure of the data. 1D CNN network has a relatively small architecture which require simple array operations. It is easier to train and test the data using 1D CNN model. Along with less computation time, 1DCNN has low hardware requirement and is best suited for real time and low-cost applications. The input to CNNs is a feature map which corresponds to the pixels in an image or words in a sentence or document, or characters in words. This feature map is scanned in CNNs one area at a time by filters, assuming that filters slide, or convolve, around the feature map.The configuration of 1DCNN is formed using various important parameters: 1. The hidden CNN layers and neurons 2. Kernel3. Max pooling layer 4. Activation functions. The major advantage of 1DCNN is that it results in low computational complexity, since the convolution is linearly weighted. The first layer is the convolution layer where the convolution of the input image with the kernel is performed. All the convoluted values are then stored in a feature map, the values filled in the feature map is obtained by moving the kernel over the input image by a specific stride value. After this step is the pooling layer or subsampling of data. This operation consists of applying some operation on the feature map and extracting some specific value for each region. The most commonly used pooling technique is max pooling which selects the maximum values in the input feature map of each step. Padding is another term which is commonly used in CNN network, in order to minimize the data loss by using the kernel. It refers to the number of pixels added to any image which is being processed by the kernel. After each convolution layer, there is a activation layer RELU, which trains the data much faster and alleviated the problem of vanishing gradient descent. The RELU layer just changes the negative values to zero and keeps the positive value as it is. After all the feature extraction done in the convolution and pooling process there comes the full connected layers for the last step where the network is trained.

III. Proposed Model

In this experiment we have proposed a 1DCNN model which takes the advantage of the multi text sentimental review dataset, with varied reviews. Preprocessing activities like removing stop words, lemmatizing and tokenizing the data is done before it is fed to the model. Cleaning of the unsampled data is performed to achieve effective results. The model built consist of all the layers of CNN network which are max pooling, convolution and dropout layers. The final output or classification of the test data is displayed using the accuracy and precision of the data. For calculating the accuracy the categorical accuracy is used and activation function used is SoftMax.

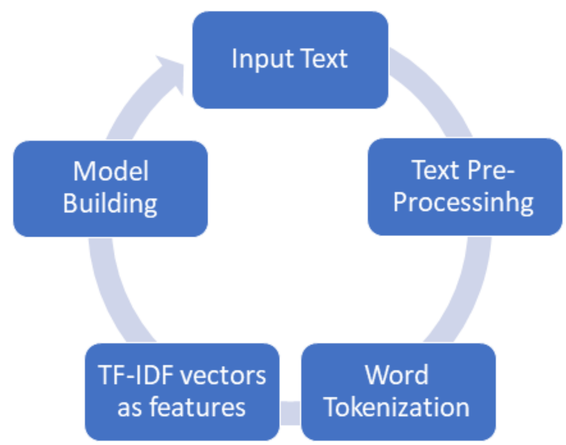

Fig. 1. Architecture of proposed model

Dataset description: The movie dataset is obtained from famous movie review website the Rotten Tomatoes, in which the sentiments varies largely. The dataset size is of 156060 reviews amongst which some are full sentences, and some are not. The main headers for the dataset are phrase ID, Phrase, sentence Id and the sentiment. Since this is a multi label dataset the sentiment value ranges from 0-4 with each phrase. Modular coding: Modular programming is the process of subdividing a computer program into separate sub-programs. It enables to work and debug the pieces of the program independently. Preprocessing: Since the dataset used for this experiment is a text-based data, some preprocessing steps are required in order to achieve a optimum result. The data consists of many sentences which do not play role in classifying the review and are not full sentences. Hence the first step is to only store full sentences and make a list of it and do the further processing on the same. After sentence extraction, data imbalance is handled by tokenizing the sentences. The further processing done on the data are removing the stop words, stemming and lemmatizing the document. The inbuilt python libraries which help in effective stemming are used for this experiment. Feature extraction: There are many approaches to effective feature extraction for textual data. The Word2vec model is a predictive deep learning model used to compute and generate high quality continuous dense vector representations of the words which have contextual meaning attached to it. In this model we can explicitly specific the size of the word embedding vectors and the total number of vectors in the corpus. From this model we can take advantage of other models like Bag of words and Term Frequency -inverse document frequency (TF-IDF). The BOW model simply represents the text as a bag of words irrespective of the grammar. This model is commonly used where frequency of words is used as a feature for the training the classifier. The usage of this model demands small vocabulary size for better results. On the other hand, TF-IDF is a numerical representation of how important the word is in the corpus. The value of TF-IDF increase as the occurrence of a particular word in the corpus increases. This helps in adjusting the judgement that some words have high weightage and can be used as important ones for training the classifier.

Splitting the data: The document obtained after preprocessing is then spilt into test and train data with a ratio of 70:30. Since all the columns of the data are not required for the processing of the result, we just consider the preprocessed text and the sentiment as the input to the model. Model building: The CNN model built is sequential and has all the required layers for processing the data. It starts with the convolution layer followed by the max pooling layer and the dropout layer. These series of layers keeps on repeating for better feature extractions and at the end the flatten layer is added. The activation function used here is SoftMax which is applied just before the output layer. This model is then fed the training data to extract the features and display the desired accuracy. The model once trained is then used to test the data to know the accuracy and precision of the model built. The final results of this experiment are displayed with thee factors like accuracy, precision and F1 score.

IV. Experimented Results

The performance of any model can be described by evaluating some of the performance metrics. The most popular metrics involve are Accuracy, precision and F1 score. These metrics help us in comparing our model to other models and also gives a clear picture of how well the model performs.

	precision	recall	f1-score	support
0	0.49406	0.31517	0.38484	2110
1	0.52043	0.53273	0.52651	8081
2	0.73085	0.76580	0.74792	23920
3	0.54218	0.54044	0.54131	9929
4	0.52211	0.42081	0.46602	2778
accuracy			0.63700	46818
macro avg	0.56193	0.51499	0.53332	46818
weighted avg	0.63146	0.63700	0.63279	46818

Fig. 2. Performance Metric Summary

The detail of the metrics is explained in the further section which gives us a good understanding of the metrics and how they are responsible. Accuracy: This is the most commonly used metric for nay model built. It is simply a ratio of the correctly predicted observations to the total number of observations. A high value for this metric gives appositive feedback about the model. But accuracy alone cannot judge a model's performance as it does not consider all the other parameters of the model.

Precision: This metric is similar to the accuracy, only that is the ratio of the correctly predicted observations to the total positive observations in the corpus. A high precision means the model has high performance.

Recall: This metric basically defines the sensitivity of any model. It gives a percentage of the correctly predicted observation to the labelled observations in the corpus.

F1 score: It can be considered as the weighted average of the above metrics defined. It considers both the positive and the negative observations.

The aim of this model proposed is to correctly classify the multi class dataset and display the result with the metrics defined above. The detailed steps involved in building the model are explained before. Based on all the steps used to clean the data and then input it to the model results in optimized result. The model is tested in various combinations of parameters.

The following table displays the performance metric of the model.

1: Epoch : 3 Batch size: 128

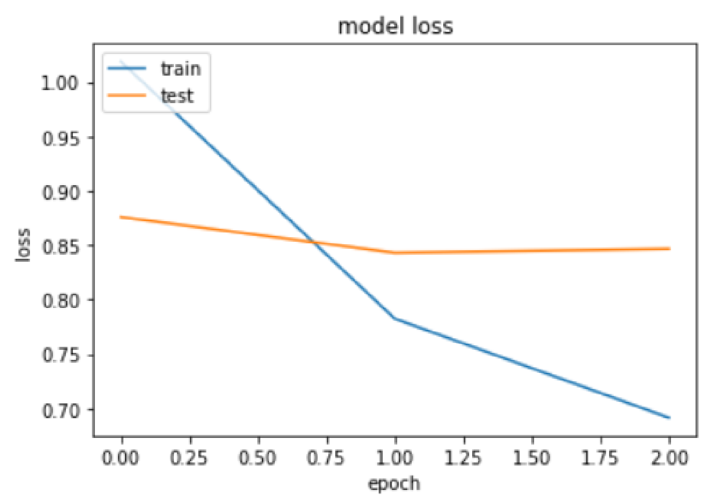

Fig. 3. Accuracy Result with 3 epoch

The next graph shows the loss for more epochs where we see the loss converging as epoch increases.

2: Epoch : 10 Batch Size:128

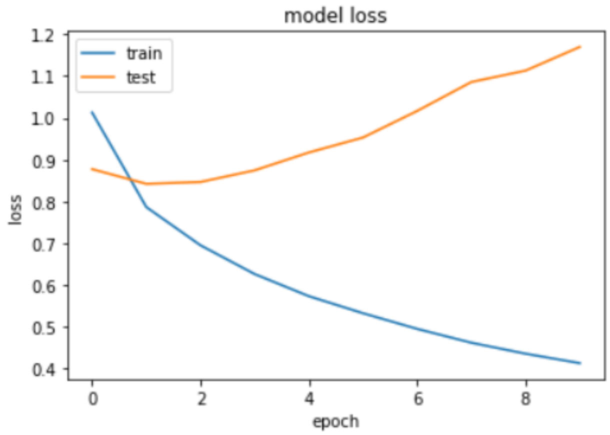

Fig. 4. Accuracy Result with 10 epoch

V. Conclusion

In this paper we have presented our systems for multi-class sentiment classification and we found that the deep neural network model can outperform traditional methods that rely on language-specific feature engineering. We show that the class imbalance in the data can lead to degradation in the system performance and point out that oversampling can be a helpful workaround for handling this imbalance. The use of various preprocessing techniques and feature extractors helped us generate a optimized input for the model which directly targeted on the reviews and all other noise was removed which could affect the performance. We show that the class imbalance in the data can lead to degradation in the system performance and point out that oversampling can be a helpful workaround for handling this imbalance. Different performance metrics are calculated to display the performance of the CNN model built.

References

[1] Attia, Mohammed, Younes Samih, Ali Elkahky, and Laura Kallmeyer. ”Multilingual multi-class sentiment classification using convolutional neural networks.” In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[2] Yasen, Mais, and Sara Tedmori. ”Movies Reviews Sentiment Analysis and Classification.” In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), pp. 860-865. IEEE, 2019.

[3] Sharma, Pallavi, and Nidhi Mishra. ”Feature level sentiment analysis on movie reviews.” In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), pp. 306-311. IEEE, 2016.