Multi-class Sentiment Analysis using Deep Learning

PANKTI RAKESHBHAI PATEL
Student Id: 1111102
MSc in Computer Science
Lakehead University
Email:ppatel63@lakeheadu.ca
Guided By: Dr. T.Akilan

***Abstract*—In the period of big data there was a massive rise in the amount of data, such as video, audio and text. The text data are the largest of all forms of data, and large-scale studies have been performed on text analysis. The research conducted focuses on one of the popular neural network modelling algorithms i.e. Convolution Neural Network. To be more precise, the paper gives a detailed research on sentiment analysis on text-based movie reviews. Moreover, multi-class sentiment analysis is been performed by using the dataset of Rotten Tomatoes (raw trained dataset) was castoff for calculating the f1 score, accuracy and many other metrices.**

***Index Terms*—Sentiment analysis, Naive Bayes, Support Vector Machine, Movie Reviews, Natural Language Processing, Machine Learning, Deep Learning, Neural Network**

I. INTRODUCTION

With the proliferation of reviews, ratings, recommendations and other forms of online expression, online opinion has turned into a kind of virtual currency for companies looking to market their products, identify new opportunities and manage their reputations. Reviews are an important means of measuring the quality of any movie. Although the quantitative distribution of a numerical rating to movie teaches about movie efficiency, a collection of movie reviews is what gives us a deeper qualitative insight into various aspects of the movie. A review of a text movie tells us about the strength and weakness of the movie, and deeper study of movie reviews will tell us if the movie meets the norm of the reviewers.Moreover to obtain semantic text knowledge, sentiment analysis is used.

A. SENTIMENT ANALYSIS

Sentiment analysis is a fundamental yet a critical problem in natural language processing. It involves multiple important theories and common techniques that are widely used in text mining. With the development of machine learning methods, there are more emerging approaches that could be applied in sentiment analysis to improve the judgment accuracy. Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics for the retrieval and recognition of subjective knowledge in source materials. It aims to decide a speaker's or a writer's attitude towards some subject or the contextual difference of a document as a whole. The attitude may be,

1) His or her opinion or decision.
2) Affective condition (that is, the author's emotional state while writing).
3) The desired emotional contact (that is, the author wishes to have an emotional impact on the reader).

B. DEEP LEARNING

The standard machine learning techniques and Neural Networks were not adequate to understand the mining and analysis of the massive data being exchanged and generated on a regular basis, hence deep learning was the key. Discussing about Deep learning, it is a method of artificial intelligence that imitates the workings of the human brain while processing data and generating patterns for use in decision making. Deep learning is a branch of Artificial Intelligence (AI) machine learning that has networks capable of learning from unstructured or unlabeled data without supervision. Most modern deep learning models are based on artificial neural networks, specifically Convolutional Neural Networks (CNN), although they can also include layer-wise propositional formulas or latent variables in deep generative models such as nodes in deep-belief networks and deep Boltzmann machines.

II. LITERATURE REVIEW

Existing approaches to evaluate emotions can be divided into four main groups. They are the spotting of keywords, lexical similarity, statistical approaches, and techniques at concept level. Keyword spotting classifies text by types of affect based on the inclusion of unambiguous words such as good, sad, fearful and bored. Lexical affinity strengthens keyword-based approach by considering terms that affect not just the obvious. It also assigns ambiguous terms to specific emotions with a possible "affinity." Statistical methods affect machine learning elements such as latent semantic analysis, support vector machines, bag of words, and semantic orientation. Unlike the above-mentioned simply syntactic methods, concept-level approaches exploit elements from the representation of knowledge such as ontologies and semantic networks so that they can also identify meanings that are articulated in a subtle way, that is by examining concepts that do not convey relevant information directly but are implicitly connected to other concepts that do so.

Stanford University has introduced a novel approach to an study of sentiments. Most traditional sentiment prediction systems operate by looking at words in isolation, offering positive points for positive words and negative points for negative words, and then summarising them. The order of words is overlooked in that method, and essential knowledge is lost. By comparison, this approach's new deep learning model actually builds up a representation of entire sentences based on the structure of the sentences. It measures the sentiment by considering how words make up the meaning of longer sentences. The model isn't as easily tricked as previous ones by using that kind of approach.

III. PROPOSED MODEL

A. DATASET

The dataset provided is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset. To be more precise, our problem description is to evaluate the sentiment of the Rotten Tomatoes dataset reviews of the movies. The dataset has been divided to training and test set with the ratio of 70:30 for the purpose of making the 1D Convolutional Neural Network model, but the sentences have been shuffled from their original order. Each phrase has a phrase Id and each sentence has a sentence Id. Phrases that are repeated (such as short/common words) are only included once in the data.The train.tsv contains the phrases and their associated sentiment labels. The test.tsv contains just phrases.

TABLE I
DETAIL ABOUT DATASET

Attribute Name	Data type	Description
PhraseID	Float64	numbering of the each sentence in incremental order.
SentenceID	Int64	ID of the unique sentence.
Phrase	Object	reviews of the dataset.
Sentiment	Int64	classification of sentiment in 0 to 4.

Sentiment label to each phrase in test file should be assigned.The sentiment labels used in the data set are,

- 0 - negative
- 1 - somewhat negative
- 2 - neutral
- 3 - somewhat positive
- 4 - positive

B. LIBRARIES USED

Defining a convolution model in order to train and test the movie review dataset. The essential libraries used in implementing this model are :

1) Keras
2) Pandas
3) Numpy
4) Random
5) Matplotlib
6) Sklearn

I have used various libraries to perform different tasks accordingly. Initially pandas are used to load the dataset. Pandas is often used with 1D-CNN to perform the classification. Mathematical calculations can be carried out using the numpy library, and it also transforms data-frame pandas into series. The CNN model is built using Keras system. The random library generates a list like a set, and then returns one element randomly from that list.

C. DATA PREPROCESSING

In order to solve the problem of overfitting, various preprocessing steps are carried out. While discussing about data preprocessing, it is a technique for data mining which transforms raw data into a comprehensible format. Real-world data are often incomplete, inconsistent and/or incomplete and likely to contain several errors. Pre-processing data is one of the best way to solve these problems. Preprocessing steps includes steps like tokenization, stemming, Stop Words Removal and Punctuation Removal and embedding. The dataset consists of sentences or a single word as a part of movie review so it must be filter before performing preprocessing steps. Therefore, to perform training and testing, whole sentences are required. Also, there is a reduction in dimension of the dataset. Now, a detailed of preprocessing steps is given below.

1) Tokenization
   Tokenization in NLP is a very simple process. This is basically a process to break a character into pieces called a token and, like punctuation, to throw away the different characters simultaneously.

2) Stop Words Removal and Punctuation Removal
   'Stop words' are commonly used words that are unlikely to have any benefit in natural language processing. These includes words such as 'a', 'the', 'is' and the method of removing these words is known as stop word removal. Removing stop words is an significant factor in classifying the text and interpreting feelings as the terms describing a sense of text can be centred more. The punctuation is omitted so that these words lack little meaning in relation with that kind of dimension.

3) Stemming
   Removal of affixes and ending of words which are based on common ones are known as stemming. This can be helpful only sometimes as often the new word is so much a root that it looses its actual meaning Any word's root is referred to as lemma. Stemming is important in the study of feelings since different terms involve different feelings.

4) Embedding
   Processing any textual data or extracting information from text involves a technique to transform strings or provided text into a real number (vector) known as Word Embedding or Word Vectorization, this method of calculating the real number from a word. The method is Term Frequency-Inverse Document Frequency (TFIDF). I used the Word Vectorization (Term Frequency-Document Frequency(TF-) technique. The justification for using TF-is that it functions in such a way that it

eliminates common-values used in different documents. There are various approaches:

1) BAG OF WORDS: We obtain the vocabulary list from the corpus (whole text dataset). The length of the vocabulary list is equal to the length of the vector that will be output when we apply Bag of Words (BOW). For each item (could be an entry, sentence, line of text), we transform the text into a frequency count in the form of a vector. We code this by setting up a count vectorizer from sklearn's library, fit it on the training data and then transform both the training and test data.

2) WORDD2VEC: Word2vec (Word to Vector) is a two-layer neural net that processes text. Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. The algorithm was created by Google in 2013. The 50-D space can be visualised by using classical projection methods (e.g. PCA) to reduce the vectors to two-dimensional data that can be plotted on a graph.

3) TF-IDF: Word Vectorization is used to map terms from vocabulary to the corresponding vector of real numbers used to classify predictions of terms or semantics of words. TF-IDF consists primarily of two concepts:

- Term Frequency(TF)
- Inverse Document Frequency(IDF)

TF means how often a word appears in a document. The mathematical expression for TF is as follows:

$$TF = \frac{Number\_of\_times\_word\_appear\_in\_the\_docs}{Total\_number\_of\_word\_in\_the\_docs}$$

IDF states the information about how to find the importance of the word. Here, the words which are less frequent are more informative and are very crucial in the document. The mathematical expression for IDF is as follows:

$$IDF = \log_{10} \frac{Number\_of\_docs}{Total\_No\_of\_docs\_in\_which\_word\_appears}$$

IV. EXPERIMENTAL ANALYSIS/COMPARISONS WITH OTHER MODEL

In this section, we will present the experiment result and show the comparison between some of them to evaluate the performance of each model. A comparison was performed by using different aspects like batch size, different optimizers for better accuracy. Moreover, a better result was gained by changing the learning rate.

A. BATCH SIZE

Batch size is a term used in machine learning which refers to the number of training examples used in a single iteration. The larger the batch size the lesser the accurate result.

TABLE II
COMPARISON OF BATCH SIZE AND ACCURACY

BATCH SIZE	LOSS	ACCURACY
256	0.993	0.6115
128	0.999	0.6127
64	1.009	0.6182

B. OPTIMIZERS

Optimizers are algorithms or techniques used to adjust the properties of the neural network, such as weights and learning speeds, in order to minimise losses.The optimizers are required to improvise the accuracy of any model.There are various optimizers like Adadelta, RMSprop, adam, sgp.I have compared the batch size with the optimizers and therafter observed the accuracy. When I allocated the batch size 64, epoch to 50 and the optimizer was Adadelta I got the highest accuracy of 62. The comparison of optimizers with respect to batch size is being displayed in Table.

TABLE III
COMPARISON OF BATCH SIZE AND ACCURACY

OPTIMIZER	ACCURACY	LEARNING RATE
ADAM	0.5837	0.01
ADADELTA	0.6182	0.01

C. EVALUATION PARAMETERS

Precision and recall are the two metrics that are commonly used to test execution in mining content and in the field of content analysis, such as data recovery. These parameters are used for calculating accuracy and completeness, respectively.

F-1 score:
The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula to calculate f-1 score is:

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{1}$$

D. RESULT

Finally, after running the code, table 4 shows the final best accuracy , f-1 score, precision, recall metrics obtained by using Adadelta optimizer.

TABLE IV
RESULT

METRICS	VALUES
ACCURACY	0.6182
F-1 SCORE	0.5879
PRECISION	0.6661
RECALL	0.5281
LOSS	0.9917

Fig.1 shows the graph of the model accuracy with respect to loss function and epoch.

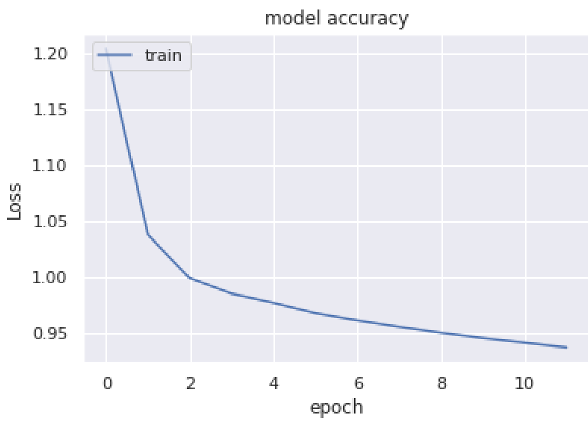

Fig. 1. LOSS VERSUS EPOCH

V. APPLICATIONS

CNN models are used to gain best results in semantic parsing, search query retrieval, prediction, classification, traditional NLP tasks and so on. Also, 1-D convolutions are use on time series in the frequency domain by unsupervised model to detect inconsistencies in the time domain.

While discussing about applications of sentiment analysis, it is a valuable tool that can be used in social media, consumer feedback and customer service for businesses. It assists in measuring public opinion about an event or product. Moreover, it is used in tracking social media and Vocabulary to track consumer feedback, survey responses, competitors etc.

Also, classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Classification is a data mining (machine learning) technique used to predict group membership for data instances.Moreover, classification is also used for speech recognition, handwriting recognition, biometric identification, document classification and so on.

VI. CONCLUSION

This assignment is based on predicting sentiment values based on the dataset of the movie review. I have used the Rotten Tomatoes movie review dataset, which consists of 5 meaning-based sentiment values and also has more than One Lakh instances to make predictions. In addition, the TF-IDF method is used to convert sentences to a matrix format. Using the Keras framework, I used the matrix to construct the CNN model and then calculate it after a precise hyperparameter tuning. Finally, I got the highest accuracy value of 61.58 percent after the hyper parameters had been changed.

REFERENCES

[1] https://www.kaggle.com/edchen/text-vectorization

[2] https://towardsdatascience.com/a-beginners-guide-to-preprocessing-text-data-using-nlp-tools-5cb52a8d3cd

[3] Attia M, Samih Y, Elkahky A, Kallmeyer L (2018) Multilingual multi-class sentiment classification using convolutional neural networks. In: chair) NCC, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T (eds)

[4] https://medium.com/@sdoshi579/optimizers-for-training-neural-network-59450d71caf6

[5] https://pythonhealthcare.org/2018/12/14/101-pre-processing-data-tokenization-stemming-and-removal-of-stop-words/

[6] https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

[7] https://www.kaggle.com/ashukr/implementation-of-word2vec-paper