Predictive Analysis using Convolution Network on Sentiment Analysis of Text Classification using Machine Learning

Vanitha kakollu[1], Radhika Pulicherla[2], G Sri Sowmy[3,] K.Naga Soujanya[4]

[1] Assistant Professor, Department of Computer Science, GIS, GITAM (Deemed to be University) [2,3,4] Assistant Professor, Department of Computer Science Engineering, GIT, GITAM (Deemed to be University)

Email: [1]vanithagitam@gmail.com, [2]radhikapulicherla@gmail.com, [3]sgudipat@gitam.edu, [4]sketha@gitam.edu

Abstract: Today we have large amounts of textual data to be processed and the procedure involved in classifying text is called natural language processing. The basic goal is to identify whether the text is positive or negative. This process is also called as opinion mining. In this paper, we consider three different data sets and perform sentiment analysis to find the test accuracy. We have three different cases- 1. If the text contains more positive data than negative data then the overall result leans towards positive. 2. If the text contains more negative data than positive data then the overall result leans towards negative. 3. In the final case the number or positive and negative data is nearly equal then we have a neutral output. For sentiment analysis we have several steps like term extraction, feature selection, sentiment classification etc. In this paper the key point of focus is on sentiment analysis by comparing the machine learning approach and lexicon-based approach and their respective accuracy loss graphs.

***Keywords*: Sentiment Analysis, Positive, Negative, CNN, Feature Selection, Machine Learning**

1. Introduction

Sentiment analysis plays a major role in today's world as it helps to analyse a person's opinion or opinion of a group of people on a particular topic. Well, this analysis can be applied in the field of business to know how well a product is doing or in the field of programming to identify user typing patterns. It is also used to predict and analyse a large amount of data on the web like attitudes and emotions. Sentiment analysis is also used to analyse pictures and videos. Most of the people express their views today through blogs or through social networks and knowing their opinions and responses of other people has become very important. Thus mining the opinions of people has at most importance today. Research in the field of machine learning has always been the topic priority for many companies. In fact, predictive analysis is ranked at 8th position compared to other 37 technologies. These machine learning algorithms and analysis techniques have helped more than 40% of the companies to grow. Analysis prediction is also one the most widely taken up project in 2019 according to Forbes, one of the leading American business magazines.

2. Relatedwork

W.Sriurai, in her paper discussed about the various machine learning algorithms and the various ways to implement the feature extraction techniques. The author has given some idea about the algorithms and techniques to use for analysis of the data like BOW.

Walaa Medhat, Ahmed Hassan, HodaKorashy have discussed a lot about the various implementations and techniques required to study the patterns of the text and classify the text accordingly. A few techniques portrayed in their paper have also been discussed here.

3. Applications

1. Decision Making: Opinions play a major role in our life. The analysis and the final answer to our analysis helps us achieve various answers to specific topics like "how was the movie", " who will win the elections"etc.

2. Trend Analysis and Predictions: Public play a major role in sentiment analysis by providing reviews which further help in predicting market trends. The expectations what people have and the requirements of the people can be found.

3. World of Business: Well, today we have a plethora of companies and the competition among these companies is also high. Every company seeks to provide top notch and innovative products. This can be done by assessing the clients or the consumers who use the product and finding the apex point on which the review depends on.

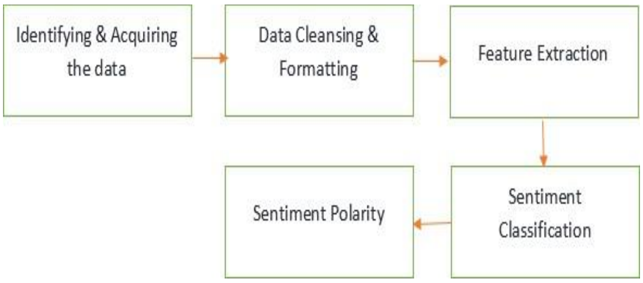

Fig. 1.Steps defining the Sentiment Polarity

4. Methodology

ISSN (Print): 2204-0595

ISSN (Online): 2203-1731

A) Choosing

Data Set: Data set is required to train the model to predict the sentiment of the sentence. And this involves selection of features.

B) Feature selection:

Selecting features is step one of sentiment analysis as it is used to train the model.

1. Stemming: Stemming is a process of removing inflection in words to reduce them to their root. Removing inflection is done through dropping of unnecessary characters, usually a suffix.

2. *Lemmatization*: It's another way to remove inflection by using detailed database of the language.

3. *N-Grams*: A combination of N words together are called N-Grams. These provide generally more information than a single word or unigram. Sentiment analysis becomes easier when we use N-grams.

4. *POS-Tags*: These tags are useful to remove stop words which are basically low frequency or not so important words. It helps to find the various parts of speech and the correlation among them and their associations with the contiguous words. We can name a few parts of speech like verbs, nouns, adjectives and prepositions.

5. *Stop Words*: Pronouns and the basic articles come under stop words. We have pronouns like he/she or them. Articles are a, an and the. It is better to remove these stop words to improve the efficiency of the program.

C) Sentiment Classification:

We basically mention two main approaches to sentiment analysis:

1. *Lexicon(Subjective):* We have a group of words which basically indicate positive or negative or neutral based on the individual words and this total collection of words is called as subjective lexicon. We have various approaches as defined below:

a)Dictionary Based Approach

b)Corpus Based Approach

2. *Machine Learning*: This is a process where the system is trained as per our requirements. Major point is text features which are converted to vectors and further processed. In this paper we are using convolution neural networks to achieve high accuracy.

D) Defining a Baseline Model: It's a simple model which acts as a base to more advanced models and helps us check how our model basically works in the initial stages. We split the data into test set and train set and have to take care of overfitting.

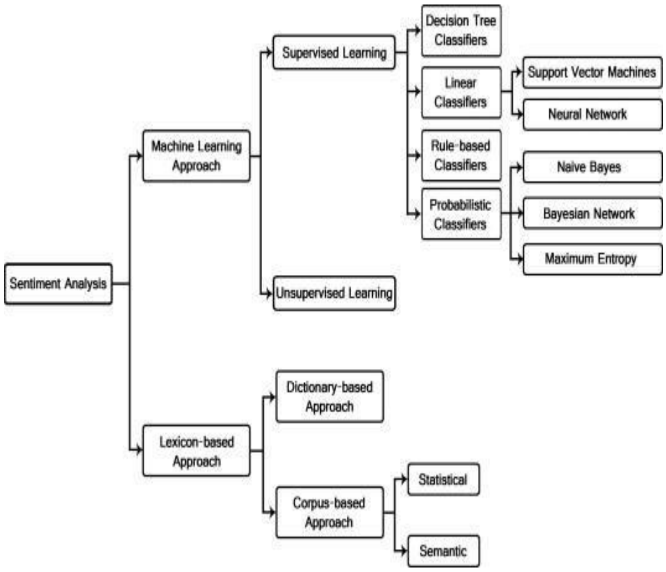

Fig. 2.Sentiment Analysis Classification

5. A Primer on Neural Networks

Neural networks have a very important role in theadvancements in the field of Artificial Intelligence. Neural networks have their wings spread across various fields like regression, classification and generative models. Mostly the fields include computer vision based classification,natural language processing, voice recognition and text prediction. These neural networks works like neurons in our brain by using a feed forward network where each layer has a set of neurons and each neuron is connected with the other in the next layer. We have an input layer an output layer and several hidden layers.

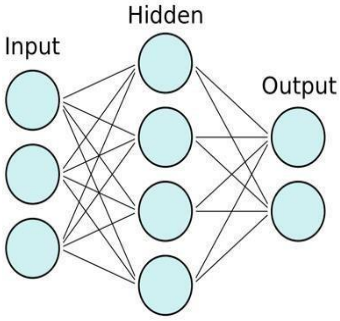

Fig. 3.Neural Networks.

We can use any number of hidden layers as we wish. To get into the mathematical depths to get an ideological visual perspective of the mathematics being used we can consider this formula which is to move from one surface to the next using this short equation:

$$O_j = f(\sum_i w_{i,j} a_i + b_i)$$

As we are moving from one layer to the other we have two layers to deal with. Here for the layer with nodes o the layer with nodes a act as the input. To get the output for each node

ISSN (Print): 2204-0595

ISSN (Online): 2203-1731

we have to add the bias b to the product of each input node by the weight w.

Now all these values are added up and passed to the function f. There are various different types of functions and these functions are basically based on the problem being dealt with or the layer. The above function is also called the activation function. We basically use ReLU (Rectified Linear Unit) to easily achieve the goal of the activation function for hidden layers.Generally neural networks basically use the concept of weights. The entire network depends on the calculation of weights and this is a very complex and difficult process. We start by taking some random numbers and keep updating the weights continuously by the concept of backpropagation.

We have to keep updating the weights and this is done with the help of optimization techniques. The error is given by a loss function and we have to work on reducing this loss. To minimize this loss we use the optimizer.Keras is a ML and neural network API which basically providesforemost building blocks for ML models. We now use a sequential model which has linear stack of layers. But the most common and widely used layer is the dense layer. Now we finally build our model which has a certain accuracy and we can use the matplotlib to obtain the graph for accuracy and loss.

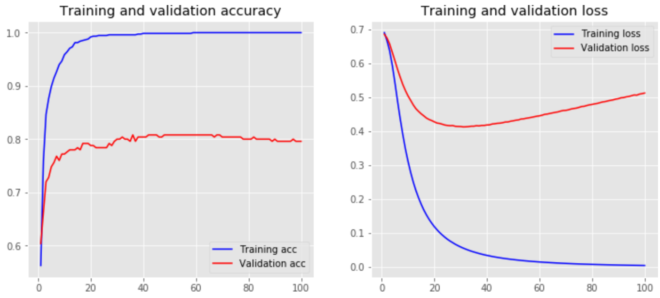

Fig. 4.Accuracy and loss graph.

We can see that training our model for too long can result in 100% accuracy with the training set this could lead to overfitting. This is for our first keras model we planned on checking.

Word embedding can be done in many ways and is important because words can be represented as vectors. This is crucial when we have large amounts of textual data. There are few parameters which are basically used by the keras embedding.

1. The size of the vocabulary and is denoted by input_dim.
2. The size of the vector(dense) and is denoted by output_dim.
3. The length of the sequence and is denoted by input_length.

Considering this word embedding concept the output for the first model is as follows: (includes keras)

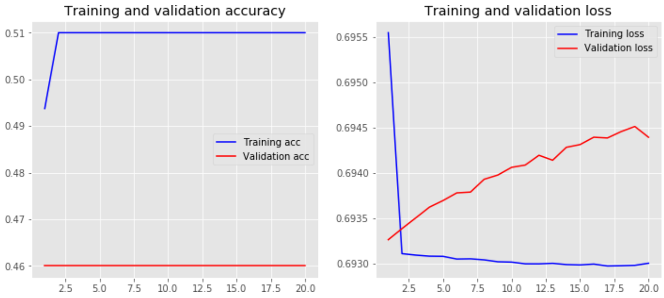

Fig. 5.Accuracy and loss graph.

But when dealing with sequential data we have to concentrate on various factors like the local and sequential information instead of completely depending on the absolute positional information. The loss and accuracy for max pooling model:

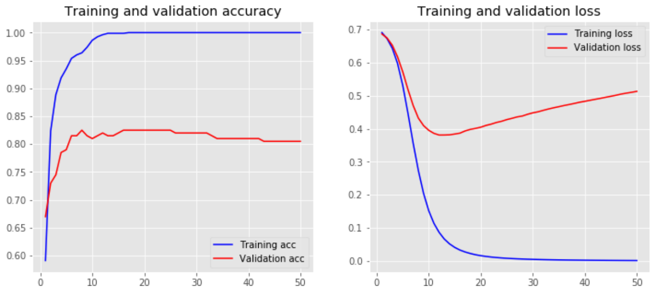

Fig. 6.Accuracy and loss graph.

We can also consider pretrained word embeddings which utilize a much larger corpus. We use two popular methods :

1)Word2Vec

2)GloVe

If we consider a large page of textual data consisting of more than 40000 words you can use word embedding as it consists of limited number of words in the vocabulary. The graph for accuracy and loss for untrained word embedding is as follows:

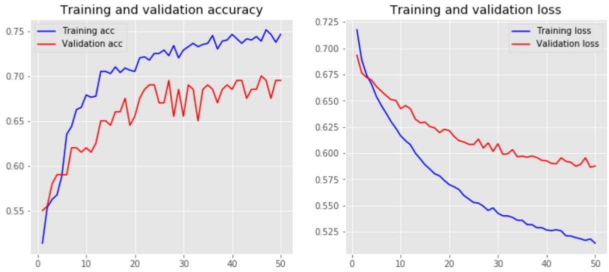

Fig. 7.Accuracy and loss graph.

If we consider the graph for accuracy and loss for pretrained word embeddings it is as follows:

ISSN (Print): 2204-0595

ISSN (Online): 2203-1731

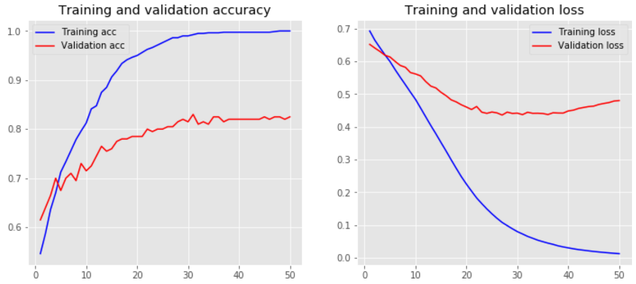

Fig. 8.Accuracy and loss graph.

6. Explanation of the Dataset and its Features

The dataset basically consists of three major class groups IMDb, Amazon and Yelp. The objective is to predict the accuracy with which one can classify whether the opinion is positive or negative or neutral. We have a large vocabulary with 4603 words in each of the class and the maximum length of each word is fixed to be 100. The kernel size is chosen to be 5.We can consider each review to have a particular score like 0 for negative comment and a 1 for the positive comment.

7. Results & Model Evaluation

Convolution Neural Networks also called as Convnets. They play a significant role in classification of images, videos. This concept can also be applied for sequential processing. These convolution neural networks use several convolution layers which consist of multiple filters which help in detecting specific features. The main heart of the technique is convolution where each layer in the group is capable to identify sparse complex patterns.While dealing with sequential data we consider a single dimension but the overall procedure remains the same.

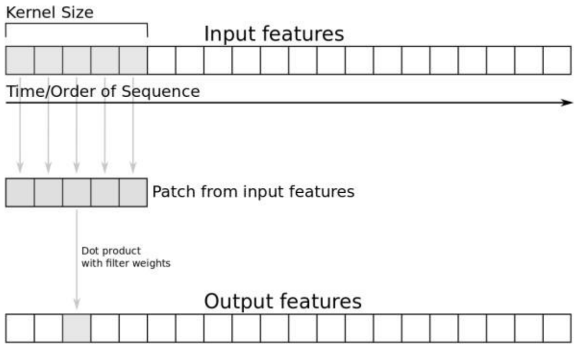

Fig. 9.Kernel with single dimension.

In the figure you can see how the how to work with a single dimension. We begin by taking the input datafeatures however its size limit must be equal to the size of the kernel filter.In the 1D convolution networks sometimes it is hard to recognise sequences at different positions.

Thekeras does offer various layers with which analysis is made much simpler. The layer what we need is a 1D-Layer and it again has several parameters to choose from. We generally now concentrate on the three basic factors or parameters like the activation function, kernel size and the number of filters.

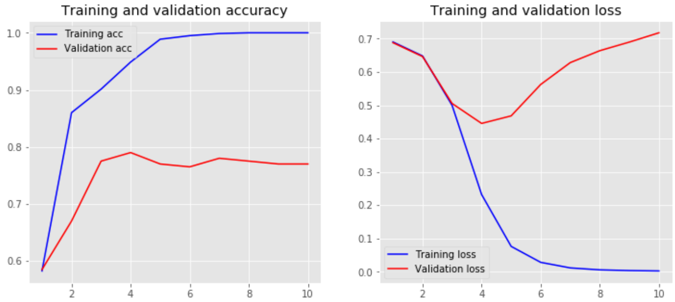

Fig. 10.Accuracy and loss graph.

The above graph gives the accuracy and loss for Convolutional Neural Network.

Finally achieved an accuracy of 80% and this seems to be a resilienthitch to get over with the information and thenetwork might not be expertly furnished.

Causes:

1)Training examples are not sufficient.
2)The data does not generalize well.
3)Hyperparameters are not tuned properly.

The major step involved is to optimize hyperparameters. We have to choose parameters which are very important and these parameters are called hyperparameters. This consumes a lot of time and there is no standard selection of these parameters.

This hyperparameter optimisation can be done with the process of grid search.There involves a lot of computational burden. It is recommended to use random search with cross-validation. It is used in majority of the problems.

8. Conclusion

We have finally learnt how to classify text using keras layers and have gone from simple methods like bag-of- wordsto more advanced and superior methods like CNN. The importance of word embeddings and how pretrained words are used in our training. We have seen the applications of hyperparameters and have achieved an accuracy rate of 80%. This convolution neural networks can be further applied to problems like spam detection, tagging of texts and categorization of news articles. Thisworkcanbeextendedtoimprovethepredictionaccuracyusinglong shot-term memory(LSTM) techniques.

ISSN (Print): 2204-0595

ISSN (Online): 2203-1731

References

[1] G. Vinodhini, "RM. Chandrasekaran Sentiment Analysis and Opinion Mining: A Survey International", *Journal of Advanced Research in Computer Science and Software Engineering.*, vol. 2, no. 6, pp. 285, June 2012, ISSN 2277 128X.

[2] G. Vinodhini, RM. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey International", *Journal of Advanced Research in Computer Science and Software Engineering.*, vol. 2, no. 6, pp. 283, June 2012, ISSN 2277 128X.

[3] W. Sriurai, "Improving text categorization by using a topic model", *Advanced Computing*, vol. 2, no. 6, pp. 21, 2011.

[4] D. D. Lewis, M. Ringuette, "A comparison of two learning algorithms for text categorization", *third annual symposium on document analysis and information retrieval*, vol. 33, pp. 81-93, 1994.

[5] Walaa Medhat, Ahmed Hassan, HodaKorashy, "Sentiment analysis algorithms and applications: A survey", *Ain Shams Engineering Journal.*, pp. 1094.

[6] G. Vinodhini, RM. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey International", *Journal of Advanced Research in Computer Science and Software Engineering.*, vol. 2, no. 6, pp. 286, June 2012, ISSN 2277 128X.

[7] Svetlana Kiritchenko, Xiaodan Zhu, Saif M. Mohammad, "Sentiment Analysis of Short Informal Texts", *Journal of Artificial Intelligence Research*, vol. 50, pp. 723-762, 2014.

[8] Walaa Medhat, Ahmed Hassan, HodaKorashy, "Sentiment analysis algorithms and applications:A survey", *Ain Shams Engineering Journal.*, pp. 1098.

[9] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features" in Machine learning ECML-98, Springer, pp. 137-142, 1998.

[10] G. S. Chanvan, S. Manjare, P. Hedge, A. Sankhe, "A Survey of Various Machine Learning Techniques for Text Classification", *IJETT*, vol. 15, pp. 288-292, 2014.

[11] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.

[12] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008;2:1–135.

[13] vanitha kakollu,” Election Result based on Twitter Data using R with Sentiment Analysis. *International Journal of Management, Technology and Engineering”, Google scholar, FEB-2020, X, 87-90*

[14] vanitha kakollu,” Identification of Co occurrence Frequencies Based on Sentiment Analysis with Spreading Activation Algorithm”. *Pramana Research Journal, Others, 2019, 9, 683-687*

[15] vanitha kakollu,” Sentiment Analysis of Election Result based on Twitter Data using R”. *International Research Journal of Engineering and Technology (IRJET), Google scholar, 2018, 5, 546-548*

ISSN (Print): 2204-0595

ISSN (Online): 2203-1731