Sentiment Analysis of Movie reviews using Neural Network

Palak Patel
(palakpatel@csus.edu)

Abstract

In sentiment analysis text is analyzed and understood and intent of text is predicted. In today's competitive world, almost every company tries to get customer's reviews in order to utilize that for marketing the product. Reviewing every comment manually will take a lot of effort so sentiment analysis can help in predicting the nature of the comment and deduce some outcome on basis of that. It is natural language processing problem. Aim of this project is to decide whether given reviews about movie is positive or negative. It is a natural language processing in which the human generated texts are classified whether that are positive or negative. I will focus on binary sentiment classifier on the given data sets to predict whether the review is positive or negative. Piece of text is examined independently or in combination to the output obtained. In order to discover the correlation, the input to neural network should be numeric and for that the text must be transformed accordingly. I will use python as programming language to build Neural Network. For coding I will use Jupyter notebook. In first phase I will focus on curating dataset, build prediction model and build neural network. In second phase I will work on reducing noise and inefficiencies from Neural Network.

Keywords: Sentiment Analysis, Neural Network, Prediction Theory,

1. Introduction

Sentiment analysis is most commonly used process to check nature of the text generated by a human to predict whether or not a section of human-generated text is positive or negative. With the help of this, a large amount of the movie reviews is obtained and is kept separately as a data resource. The process of sentiment analysis works on the basic concept of mining the opinion of viewers in order to distinguish the result into positive or negative. This kind of analysis can take into consideration the natural processing.

This process is of great importance since it can help in fetching the opinion of users and then using the data for deducing some result. Through the criterion of sentiment analysis, one can predict from the text whether the writer's opinion is positive or negative by finding the actual motive behind the text. In recent years, businesses are trying to implement this analysis in order to get the hidden motive behind the reviews. Similarly, the sentiment of movie reviews helps in better understanding of the popularity of the movies. For instance, if the producer of the movie wants to know about the public review for a given movie then the applying the sentiment analysis one can get the opinion of different people.

Sentiment analysis is a process to check whether or not a section of human generated text is positive or negative. With the increasing number of data and reviews of movie, a vast resource of data is created, which can be used for Sentiment Analysis. For example entire crew and cast of a movie would want to know the public response for their movies. Sentiment Analysis is one of the most important part of data mining. In which data can be mined based upon the positive or negative sense of given data. It is also popularly known as Opinion mining refers to the use of natural language processing, text analysis and computation of language to identify and extract subjective information from source data.

Over the past few years, sentiment analysis becomes popular because of the fact that there has been the tremendous amount of reviews, opinions of different people which act as one of the most important information to predict whether the given product will be successful or not. From this kind of data one can analyze some pattern to predict the likes and dislikes of the consumers and this can prove out to be a boon for the industry as they can predict the success rate of a given product and can modify their product accordingly.

1.1 Neural Network :

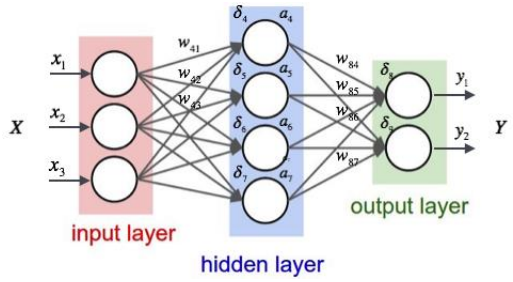
It is a supervised learning algorithm. The model learns a function by training on dataset.

$$f(\cdot) : R^m \rightarrow R^o$$

Representation of Inputs and Outputs for neural network model

In this model, input will be number of dimensions(m) and output will be number of dimensions(o). If the features are X and target y, it can learn non-linear function for either classification or regression. The difference between neural network and logistic regression is such that between the input and output layers, there can be more non-linear layers known as hidden layers.

Neural Network can be successful in discovering correlation between input and output data. Neural Network don't naturally accepts texts as input, they accepts numbers. Transform text data into numerical form in such a way that the Neural Networks can easily discover the correlation. Using NLP, statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit.



This process usually involves two steps:

• Subjectivity classification of a sentence into one of two classes: objective and subjective

• Sentiment classification of subjective sentences into two classes: positive and negative

2. Problem Formulation

The problem addressed here in the project is of analyzing the writer's opinion by focusing on the text field and distinguishing them into two different categories. The process of finding the dataset required for the document is trivial in nature. The whole project takes the chosen dataset as a base, so it must be taken into account that all the required columns that can contribute towards the desired goal must be present in the given dataset.

The dataset to be chosen must not be unstructured, noisy, unformatted, inconsistent, etc. in nature. The actual data chosen for the project initially consists of userID, type, Review, and label. The field userID is unique for different users, type field tells whether the data is test or train and the label fields indicate the reviews i.e. positive or negative.

In order to use the given dataset to get the desired results the data must be filtered, and unwanted and unstructured comments must be removed. After cleaning the data set all the different prediction algorithms mentioned above are performed on the final cleaned dataset. The output will be generated after the application of these models on the dataset and then decision can be made whether the review is positive or negative

3. Related Work

In sentiment analysis (the sub-field of natural processing) lots of research work is being performed over few years and the research is still in process. The algorithm used so far includes the Naïve Bayes, KNN. The work is being done to implement Neural Network since it provides us with the efficient results.

In most of the scenarios the scientists consider keywords as a feature to extract the sentiment from the given text field. Keywords basically includes all the different adjectives, synonyms, etc. which can ultimately guide to a common conclusion.

One of the researchers named Janice M. Weibe once performed classification on sentences. In his research work, he collected all the data from multiple sectors such as banks, movies, travel, etc. He then applied various methods to obtain the classification of the review into two different categories i.e. positive or negative. Different researchers who worked in this field has used KNN algorithm, Naïve Bayes Algorithm. The accuracy achieved was 61.81% and 80.12% respectively. In this project, we have used neural networks in addition to all the models mentioned above in order to get the accuracy of around 87%.

4. Data Set

Searching for a proper dataset was trivial in this project. Expected information from the dataset was:

DataSet should have sufficient and necessary columns to form a composite decision parameter. Using those parameters results can be obtained. It should not have a high frequency of conflicting data. It should be in an accessible and compatible format on which data preprocessing could be performed.

4.1 Data Issues

4.1.1 Noisy Data

In text if specific fields that contain unfamiliar data which is unable to understand and interpret correctly by machines for example, Unstructured text. In my dataset, the column “review” had many fields with improper structure. For example in some reviews, intent behind the review is very unclear to understand. To solve this problem first have to divide review text and separate each word and then decide based upon those words whether given review is positive or negative..

4.1.2 Unformatted Text

Some datatypes are Incompatible(Unformatted). Some of the data is in string format with punctuation mark, which affects significantly on the intent of review. They were in different formats which had to be handled while preprocessing.

4.1.3 Inconsistent Data

Sometimes in dataset there is a lack of compatibility or similarity between two or more facts(Containing discrepancies). Frequency of data is very high in all the fields where one fact can be represented in various ways using abbreviation, code names, symbols etc.

4.1.4 Data Quality

Some data contains aggregate data, lack attributes values, certain attribute of interest. Some field values are used for decision making parameters that can be missing, which affects most on sentiment to be predict. Because of that we have to add some data and there is another issue of aggregate data.

4.1.5 Performance

Performance is also known as Deteriorate without pre-processing, containing errors and outliers. If the data is inaccurate, it is not possible to achieve the expected accuracy without removing errors and outliers. This is one of the major aspects to consider to obtain efficient results. To improve performance data must be accurate and there must be ground-truth in dataset.

4.1.6 Data Skewness

Skewness is also known as measure of symmetry, or more precisely and the lack of symmetry. Dataset or distribution is symmetric if it looks the same to the left and right of the center point.

4.2 Project Data

Here in given data set it consists of 50,000 IMDB movie reviews, which are selected for sentiment analysis. Sentiment of reviews is binary classification, if IMDB movie rating<5, means sentiment score of 0, and rating>=7 have a sentiment score of 1. Individual movie has less than or equal to 30 reviews. The 25,000 review labeled training set does not include any of the same movies as the 25,000 review test set. In addition, there are another 50,000 IMDB reviews provided without any rating labels for validation cases.

labeledTrainData - The labeled training set. This file is tab-delimited and has a header row followed by 25,000 rows containing an id, sentiment, and text for each review.

testData - The test set. The tab-delimited file has a header row followed by 25,000 rows containing an id and text for each review. Your task is to predict the sentiment for each one.

unlabeledTrainData - An extra training set with no labels. The tab-delimited file has a header row followed by 50,000 rows containing an id and text for each review.

sampleSubmission - A comma-delimited sample submission file in the correct format.

5. Data Pre-Processing

Data cleaning is performed on raw using type checking and normalization. Above Data Issues are handled step by step to make sure data is consistent and compatible with the Machine Learning Algorithm.

Noisy Data is handled by filtering out the unstructured text followed by changing all the values of those in proper format.

Unformatted Text: Deciding the proper format of all the fields and changing all the unformatted values into an appropriate format.

Inconsistent Data: If some data was found to be erroneous, all other values in the respective column were considered to evaluate the mean, w hich was then entered in place of the erroneous data.

5.1 Curate a Dataset

Neural Network does is search for direct or indirect correlation between two datasets. Here I have two datasets review.txt and labels.txt, Neural Network will take

reviews as input and then try to predict whether its positive or negative. If we were working from raw data, where we didn't know it was all lower case, we would want to add a step here to convert it. That's so we treat different variations of the same word

5.2 Predictive Theory

First take a look at dataset. Try to figure out whether given text is positive or negative. Most reviews have nuance, they have particular choice of words and sequence that's not really going to be duplicated very often.

Examine all the reviews. For each word in a positive review, increase the count for that word in both your positive counter and the total words counter; likewise, for each word in a negative review, increase the count for that word in both your negative counter and the total words counter.

Common words like "the" appear very often in both positive and negative reviews. Instead of finding the most common words in positive or negative reviews, what we can do is the words found in positive reviews more often than in negative reviews, and vice versa. To accomplish this, calculate the **ratios** of word usage between positive and negative reviews.

Neural Network can't really do anything. All a neural network really does is indirect correlation between two datasets. So as network trained it needs two datasets. So eventually it can take one and learn to predict the other. So predictive theory is mainly correlation between review dataset and label dataset. For example " This movie is terrible" is consider as negative review. Most reviews have nuance, they have particular choice of words. In negative example words like terrible, horrible.

Count words both from positive and negative reviews. For that create counter which is just like dictionary. Words as input to Neural Network in such a way that it can look for correlation positive or negative prediction of output.

5.3 Transforming text into numbers

Want to present words as input into the neural network in such a way that it can look for correlation and make correct positive or negative prediction of the output.
So simply starts with,

- Count each words.
- Input those counts as input to the Neural network.

Here represent positiveness and negativeness as a number, where positive is number 1 and negativeness is number 0. Take input and output data and transfer them into appropriate 1-0 binary representation.

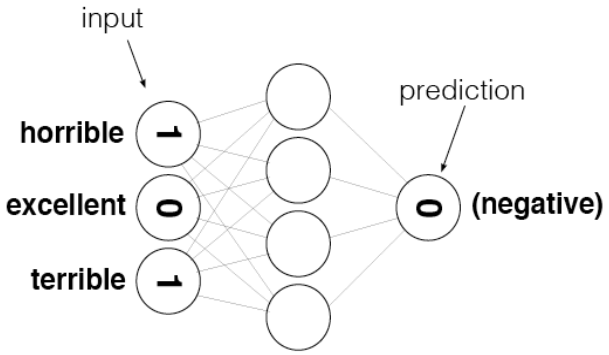


Fig 4.1 : Three layer Neural Network

6. Model Development

Building a Neural network :

- Build neural network to train dataset.
- 3 Layer Neural Network
- No non-linearity in hidden layer
- Use neural function to create the training data
- Create a "pre_process_data" function to create vocabulary for training data generating functions.
- Modify "train" to train over the entire corpus.

Going to build first Neural Network over dataset that already created.

Here I represent positiveness as 1 and negativeness as 0. Prediction is mutually exclusive, means a reviews can not be both positive and negative. Count all words that happen in a review and put them into a fixed size vector, create empty vector and added as you go, so it allocate memory dynamically and we don't have to create vector from scratch every time.

Train neural network on dataset created. Here I created Pre_Process_Data function which find length of reviews. Split each reviews and store words in dictionary. First check number of training reviews and number of training labels. Generate input data first, give data to input layer which is layer_0.

Take three layer Neural Network and remove Non-Linearity from hidden layers. Everything is self contained in class and modify the train variable to actually train over

the entire corpus. This neural Network will counts of words that are inside of review.

Create pre process data function, which is based on hundreds of reviews and unique vocabulary. Number of training review is same as number off labels. Create matrix that keeps track of positive and negative reviews.

There are two types of data propagation in Neural Network:

Forward Propagation :

- Generate input data first.
- Generate Hidden Layers
- Generate output layer without non-linearity

Back Propagation :

- Output error
- Back Propagated error
- Updates the weights

Here I select first 24000 reviews to train model and last 1000 reviews for test data. Evaluate model before training just to check results.

Under Shooting : Network was trained very slowly but it does tend to make progress.

7. Reduce Neural Noise and inefficiencies in Neural Network

Neuronal noise or neural noise refers to the random intrinsic electrical fluctuations within neuronal networks. These fluctuations are not associated with encoding a response to internal or external stimuli and can be from one to two orders of magnitude.[1] Most noise commonly occurs below a voltage-threshold that is needed for an action potential to occur, but sometimes it can be present in the form of an action potential.

Vector is list of weights, this number affects how dominantly these weights control this hidden layer. First take set of values, re-weight them and run a function, then re-weight them again do a function and then it will be final prediction.

This is a 3 steps process, which are as follow.

7.1 Understanding Neural Noise (Making Learning Faster by Reducing Noise)

Neural Noise vs Signal :
Weights control how dominantly this input affects this hidden layers. Waiting is causing it to have a dominant effect in the hidden layer and the hidden layer is all of the output layer gets to use to try to make a prediction.

If hidden layer does not have rich information then output layer is going to struggle. Re-Weight hidden layer and do a function and that's the prediction.

Reduce Noise : Remove all words that aren't relevant and listen more alternatively to the words that are relevant. If that vocab term exist then set it to 0. We can reduce noise by getting rid of this weighting. And after that neural network will be able to find co-relation so much faster. Eliminate noise by getting rid of weighting and neural network was able to find correlation much faster. It increase signal and reduce noise.

7.2 Analyzing Inefficiencies in our Network (Making Network train and run faster)

To gain accuracy and speed of training neural network we must remove inefficiencies from neural network.

Analyzing inefficiencies in Network :
In matrix representation of neural network, -- Input corresponds to rows in matrix and output corresponds to column.

If we multiply any weight with 0, it doesn't change layer 1 from what it was before. This is the biggest source of inefficiencies in network.

7.3 Further Noise Reduction (Reducing noise by Strategically reducing the vocabulary)

Here, carve out a little bit of noise so neural network can better see signal. Reduce noise in in specific ways according to different cut-offs and thresholds.

Polarity cut-off shows that positive to negative ratio has to be greater than or equal to the positive polarity cut-off or less than or equal to the negative polarity cut-off. Word has to either be less than the negative ratio or greater than positive ratio.

8. Experimental Results

After building Neural Network and reducing noise from Neural Network, we will increase accuracy.

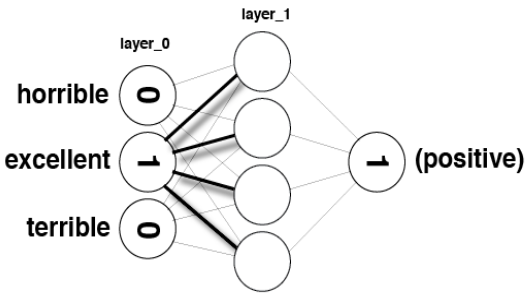
	Training Accuracy	Testing Accuracy	Speed(Reviews/Second)
Basic Neural Network	50.05%	49%	219
After Reducing Neural Noise	83.7%	85.7%	160
After reducing inefficiencies	84.5%	84.6%	1623
Further noise reduction	85.6%	87.9%	6423

With neural network we can gain only accuracy of around 50% and speed of around 200 Reviews/Second.

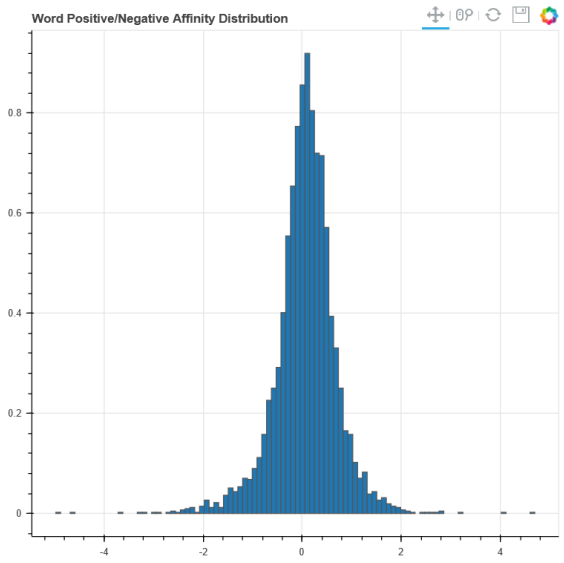
After understanding neural noise and reducing it we can get accuracy of around 82% and speed of only 160 Reviews/Second.

After reducing inefficiencies from neural network we can improve speed and get speed of around 1600 reviews/second. But we can not improve more accuracy.

To improve both reviews and speed we have reduce more noise. So after further noise reduction we can get accuracy of around 87% and speed of 6000 reviews/second.



Polarity cut-off shows that positive to negative ratio has to be greater than or equal to the positive polarity cut-off or less than or equal to the negative polarity cut-off. Word has to either be less than the negative ratio or greater than positive ratio.



9. Conclusion

Problem addressed in this project is to use deep learning can be applied to analyze sentiment of Movie reviews and for that I implemented Neural Network. Since this type of Neural Network has been showing good results in previous research. The aim of study is to increase the performance for sentiment classification in terms of accuracy and speed. First started by curating dataset and developing predictive theory. So that neural network would be able to identify correlation between the input and output data. We can validate this theory using simple count based heuristic and found we were able to identify words with both positive and negative correlation to output data. However when we train first neural network on this data, it was only barely able to identify correlation, struggling to cut through the noise, so what we did next we increase the amount of and negative correlation to output data. However when we train first neural network on this data, it was only barely able to identify correlation, struggling to cut through the noise, so what we did next we increase the amount of signal and decrease the amount of noise. First Neural Network struggle to convert all, barely reaching accuracy of 60% and speed of 100 reviews/second. Final network was able to classify 90% accuracy and speed of 7000 reviews/seconds.

10. References

[1] Andy Bromberg(2013) Second Try. Sentiment Analysis in Python. Retrieved from http://andybromberg.com/sentiment-analysis-python/

[2] . Vik Paruchuri(2015) Using Naïve Bayes to Predict Movie Review Sentiment. Retrieved from http://blogdataquestion/blog/naive-bayes-movies/

[3] Sentiment Analysis With Convolutional Neural Networks - Classifying sentiment in Swedish reviews http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1105494&dswid=-1447

[4] How to clean Text for Machine Learning with Python using nltk library. https://machinelearningmastery.com/clean-text-machine-learning-python/

[5] . Vik Paruchuri(2015) Using Naïve Bayes to Predict Movie Review Sentiment. Retrieved from http://blogdataquestion/blog/naive-bayes-movies/

[6] Dataset https://www.kaggle.com/utathya/sentiment-analysis-of-imdb-reviews/data

[7] The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers https://arxiv.org/ftp/arxiv/papers/1612/1612.01556.pdf

[8] Research Paper on Basic of Artificial Neural Network http://www.ijritcc.org/download/Research%20Paper%20on%20Basic%20of%20Artificial%20Neural%20Network.pdf

[9] Research on Sentiment Analysis : The first decade http://sentic.net/sentire2016ahlgren.pdf

[10] Analytical mapping of opinion mining and sentiment analysis research during 2000–2015 R. Piryani a , D. Madhavi b , V.K. Singhc, http://sentic.net/scientometrics-of-sentiment-analysis-research.pdf