Pune Institute of Computer Technology
Dhankawadi, Pune

A SEMINAR REPORT
ON

SENTIMENT ANALYSIS USING MACHINE LEARNING TECHNIQUES

SUBMITTED BY

ABDUL DALVI
Roll No. 31201
Class TE 2

Under the guidance of
Prof. S.S.SHEVTEKAR

DEPARTMENT OF COMPUTER ENGINEERING
Academic Year 2019-20

DEPARTMENT OF COMPUTER ENGINEERING

Pune Institute of Computer Technology

Dhankawadi, Pune-43

CERTIFICATE

This is to certify that the Seminar report entitled

"Sentiment Analysis Using Machine Learning"

Submitted by

Abdul Dalvi Roll No. 31201

has satisfactorily completed a seminar report under the guidance of Prof. S.S.Shevtekar towards the partial fulfillment of third year Computer Engineering Semester II, Academic Year 2019-20 of Savitribai Phule Pune University.

Prof. S.S.SHEVTEKAR
Internal Guide

Prof. M.S.Takalikar
Head
Department of Computer Engineering

Place:
Date:

ACKNOWLEDGEMENT

I sincerely thank our Seminar Coordinator Prof. B.D.Zope and Head of Department Prof. M.S.Takalikar for their support.

I also sincerely convey my gratitude to my guide Prof. S.S.SHEVTEKAR, Department of Computer Engineering for her constant support, providing all the help, motivation and encouragement from beginning till end to make this seminar a grand success.

Contents

List of Tables

List of Figures

Abstract

Sentiment analysis deals with identifying and classifying the input source text based on sentiments and opinions. In the business field the opinions of the customers are very valuable to companies. Customers use social media platforms like twitter, or blogging web sites to express their views and opinions. Through such platforms a large amount of data is generated. This data is used as input and analysis is performed on it.

Twitter is one of the most used social networking platform, where it allows only 140 characters at a time which has a great amount of slang words, emoticons and misspellings occurrences, hence analysis becomes a difficult task. Knowledge base approach and Machine Learning approach are the two strategies used for analyzing the sentiments from texts.

We will try to emphasize on the Machine learning approach for sentiment analysis.

Keywords

Twitter, Sentiment Analysis, Machine learning.

1 INTRODUCTION

Sentiment analysis or opinion mining is the technique of analysis text to find the opinion, attitude, emotions or appraisals of an entity like product, services, organization or issues using computational techniques like Knowledge based approach or Machine learning approach. For any computational process data is required first. In the age of Internet, people express their opinions using social platforms like blog posts, forums, online discussions, product reviews, and social networking websites like twitter. The amount of user generated data by such platforms is huge and a challenging task for humans to perform analysis on. This tedious task can be done much efficiently by computational power, thereby automating it.

The beliefs and perception of majority of people or even organizations choices or decisions, are to a considerable degree influenced by the how the masses operating on internet platforms evaluate or see it. For example, before purchasing a product, one is no longer limited to only friends and family but can seek reviews of the product given by several users through blog posts, online forums and online discussions. For an organization it may be no longer necessary to conduct surveys to gather public opinions due to the abundance of the publicly available information. Using tools like Sentiment analysis helps organizations take steps necessary to improve their product and services to attract more customers. This can be used to study the competitor's business as well

Sentiment analysis can be mainly performed using two techniques: Knowledge Based approach and Machine Learning approach. Knowledge based approach requires a large predefined dataset of emotions and an efficient knowledge or data representation for identifying sentiments. Producing such a large database of predefined sentiments can be a challenging task. Machine learning techniques don't require such a predefined database and hence proves to be much simpler than Knowledge based approach.

Machine Learning techniques make use of a training set and test set for classification. The training set consists of input feature vectors and their corresponding class label. Using this set, a model is developed which maps the feature vector to its corresponding class label. The test set is used to validate the model with unseen features to predict their corresponding class labels. Machine Learning techniques like Naïve Bayes, Maximum Entropy, Recurring Neural Network, Support Vector Machines are used as classification model. Features like Part of speech, n-grams, Term presence and Term frequency can be used for classification. These features can be used to find the semantic orientation of the document as a whole. This orientation can be either positive or negative.

Social media platforms like twitter, blog posts, online discussions and forums generate a large amount of sentiment data. This data generated by social platforms could hold reviews or sentiments of any entity like a product, service, organization or even some famous personality. Sentiment analysis of tweets is a challenging task as tweets are short in length, 140 characters at a time to be precise which occur with misspelling, slangs, and use of emoticons. Tweets are short, noisy and covers a variety of topics. Tweeters often used different vocabularies also. All of this puts a challenge to sentiment analysis.

2 MOTIVATION

Sentiment analysis has become an important tool in today's Internet information age as it holds a large amount of data. If this data could be analysed efficiently, it could promote the development of a series of applications such as automatic analysis decision making, network public opinion analysis, emergency warning or commodity sales. It has great scientific values and practical application prospects.

Twitter data or social networking sites are to know about behaviour and opinion of users. It may help to track the interest and connectivity of users with respective viewpoint of any entity like product, service or organization. Twitter is a social media platform which is widely used by individuals to large organizations who play a major role in the society. The general masses or public as well share their interests towards these individuals or organizations through Twitter. The analysis of their opinions or sentiment is held worthy for economic, social and political reasons. Sentiment analysis has diverse applications in the commercial as well as scientific field. With the internet being used to a great extent it produces data which could be used by companies to make suitable decisions or choices to maximize profits. For example, customer's reviews are of utmost important to the company to improve their services and maximize profits.

Sentiment analysis can be used to quickly gain insights on large volume of data. It used in the stock market as well to monitor a company's performance through news articles. This gives a significant financial opportunity to people to buy more stocks of a company. With access to this type of data, it gives the trades an opportunity to make efficient decisions.

Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. Thus, sentiment analysis helps in making decisions in maximum interests of the parties.

3 LITERATURE SURVEY

The Following table shows the literature survey by comparing techniques propose in various references:

Table 1: Literature survey

No.	Title	Description	Findings	Dataset
1	Sentiment analyis in Twitter using Machine learning techniques	Breif study of various methods used for sentiment analysis	Machine learning techniques prove to be much efficient than symbolic techniques	Own dataset created
2	Comparative Analysis of sentiment orientation using SVM and Naive Bayes techniques	Sentiment analysis was performed on a movie dataset using both techniques	It was found, that drama genre had the highest accuracy rate	IMDb dataset
3	Sentiment analysis on Movie reviews using recurrent neural network	Breif study of neural network	RNN has better accuracy than SVM and Naive bayes classifiers	IMDb datase
4	Deep learning for Sentiment analysis	Survey of deep learning	Overview of deep learning, applications of sentiment analysis	-

4 SURVEY ON PAPERS

4.1 Sentiment Analysis in Twitter using Machine Learning Techniques

This paper is elaborates the techniques used for sentiment analysis of tweets and research made by various researchers .Section 1 of this paper briefly talks about sentiment analysis and the two most common methods. First, is the Symbolic or the Knowledge based approach. This approach makes use of available lexicon resource. In this technique to determine the overall sentiment, sentiments of every word is determined and those values are combined with some aggregation functions. Methods followed by researchers were demonstrated as well

The second method used was Machine learning techniques. Machine Learning techniques use a training set and a test set for classification.Training set contains input feature vectors and their corresponding class labels. Using this training set, a classification model is developed which tries to classify the input feature vectors into corresponding class labels. Then a test set is used to validate the model by predicting the class labels of unseen feature vectors. Various classifiers like Naive Bayes, Support Vector Machine, Entropy Classifier and Ensemble classifier were used.

4.2 Sentiment Analysis on Movie Reviews using Recurrent Neural Network

This paper talks about sentiment analysis of movie reviews using Recurrent neural networks. The database used was IMDb dataset was used. This technique of sentiment analysis proves to be more efficient than Naive Bayes and Support Vector machines. Before actually classifying the tweets into positive negative or neutral, pre processing on data must be done.

Recurrent Neural Networks are better than traditional neural networks, as they have a memory which stores the computations of previously made calculations and make use of them for better produced outputs. RNN's have three layers : inner, hidden, and output layer. The hidden layer can further have several layers.For this sentiment analysis task, Keras1 has been used for modelling the DL2 (deep learning) models. Keras is a programming framework for deep learning and its written in Python programming language. Keras is run on top of Tensor flow which is allows fast mathematical computations. TensorFlow has tools to support Reinforcement learning.

4.3 Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques

In this paper sentiment orientaion considering positive and negative sentiments using flim reviews by user. Naive Bayes classifier technique was applied. Machine learning algorithms like naive bayes, linear support vector machines were used.

The paper is distributed in different sections. Section II discuss about the related studies about the user reviews, techniques of mining and sentimental analysis. Section III describe about the methodology of the proposed work, dataset and techniques used. Section IV discuss about the experiment that performed in this paper and its subsequent result. Section V describe the conclusion and future scope of the paper.

Before actually performing the calculations, data is collected and pre processing is performed on it. Two classification algorithms are used i.e. Linear SVM and Naïve Bayes on the movie dataset. A model is trained to evaluate the performance. The result is obtained and describes the accuracy rate of the different genre of the movie and display the end result. It was found that the movie drama has the high accuracy rate among the different genre of the movies.

5 PROBLEM DEFINITION AND SCOPE

5.1 Problem Definition

To study Sentiment Analysis of tweets using Machine learning techniques

5.2 Scope

Sentiment analysis can be used for diverse applications in various fields to maximize interests or profit of companies. It determines the opinion held of an entity with use of the large amount of data produced on the internet.
Sentiment analysis could be performed using various Machine learning techniques which uses models such as Naïve Bayes classifier, Support Vector Machine, Recurring neural networks. Each have their own advantages and disadvantages proving it to be efficient in different fields.

6 DIFFERENT MACHINE LEARNING ALGORITHM

6.1 Naïve Bayes

Naïve Bayes classifiers are a collection of algorithms base on Bayes theorem. It is family of algorithms where all of them share a common principle which is every pair of features being classified is independent on each other. The Naïve Bayes makes assumptions that each feature represented is independent and equal. The limitation of this assumption is that in the practical world it is almost impossible to get a set of independent features.

The Naive Bayes uses Bayes theorem Naïve Bayes classifier makes use of all the features present in the feature vector and analyses them individually as they are completed independent of each other, as stated in its assumption. Independent features could include emoticons, keywords, count of positive and negative hashtags which can be used for classification. Naïve Bayes does not consider the relationship between features so it cannot utilize the relationship between part of speech tag, emotional keyword and negation. It works well on large datasets. Apart from sentiment analysis these algorithms are used in spam filtering, recommendation systems as well. They are fast and easy to implement but their major drawback is the requirement of having and independent set of features which in practical applications is close to impossible. There are three types of Naïve Bayes classifier: Multinomial Naive Bayes, Bernoulli Naïve Bayes, Gaussian Naïve Bayes

6.2 Support Vector Machine

This method is a supervised machine learning technique that uses classification algorithms for two groups classification problem. It is a binary classifier and makes use of a hyperplane to separate tweets on either side of the plane. This hyperplane is used to classify data points, tweets over here. Data points falling on either side of the plane can be attributed to different classes.
The support vector machine is give a set of labelled training data of the two categories and is trained on this training data which is tested against new data which has not been used in the training set for verification purpose, this data is called the test data.

The hyperplane plane produced by the support vector machine is a decision boundary. The dimension of the hyperplane depends upon the number of input features. For example, if the input consists of only of two features then the hyperplane will be simply a line, if it consists three features then a two-dimensional plane is produced. When the features exceed three dimensions it becomes difficult to image such a plane. But in the practical world, problems with more than three features arise.

6.3 Recurring neural network

Recurring neural network (RNN) are popular and efficient models which have proven to be useful in Natural language processing (NLP). RNN make use of sequential information. The difference between a traditional neural network and recurring neural network is that the inputs and outputs are independent of each other in traditional neural networks. But in recurring neural networks the output is dependent on the

previous computational with the same task being performed for every element in sequence. RNN's have a memory which stores about the computations made so far. RNN's have three layers : input layer, hidden layer and output layer. Processing of all the input data is done in hidden layer. The hidden layer can be more than one layer, depending on the complexity of the problem. As compared to other machine learning techniques this provides high accuracy and polarity compared to other classifiers, in sentiment analysis. For sentiment analysis using RNN, Kesar1 could be used. Kesar1 is a programming framework in Python which is used for modelling the DL2(deep learning) models. This is run on top of TensorFlow which is used for mathematical computation in python and has tools to support reinforcement learning.

7 METHODOLOGY

7.1 Workflow

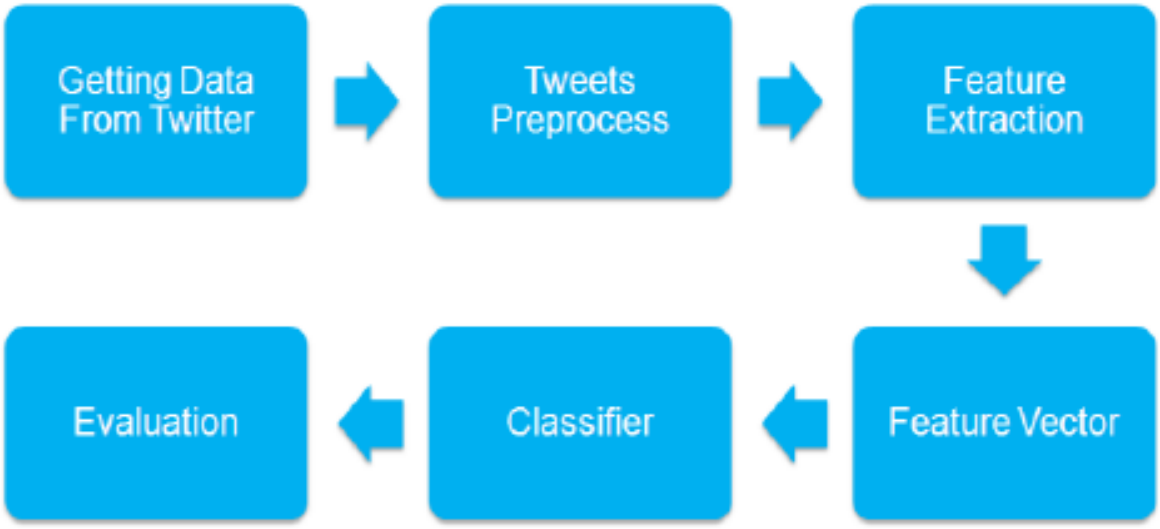

Figure 1: Workflow

7.2 Mathematical model

- Naive Bayes Theorem: This uses Bayes theorem as its base, which is :

$$P(y \mid X) = \frac{P(X \mid y)P(y)}{P(X)}$$

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

- Support Vector Machine: This helps in classifying the data into two distinct classes. It is used to calculate the best hyperplane for the given data points.

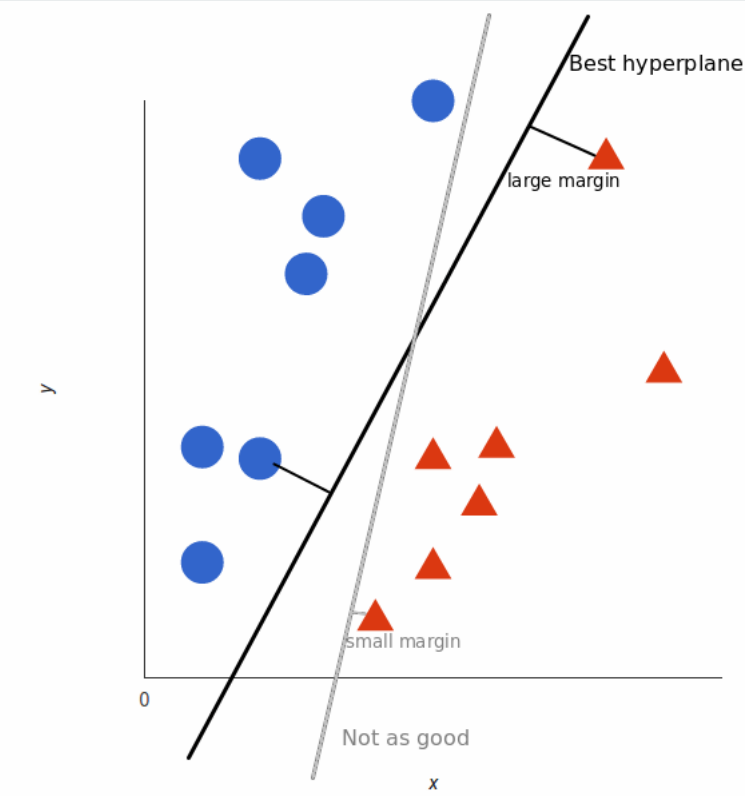

Figure 2: Best hyperlane calculated using SVM

8 Results

8.1 Data

Table 2: Data Table

S.No	Data set	size
1	Neik Sanders corpus file	232 KB

8.2 Implementation Results

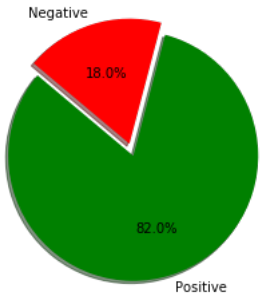

Figure 3: Result of search word "google"

In [49]:

```
1 output_list
```

Out[49]:

	Tweet	Time	Result	Search_Term	Score
0	RT @MikeParentLEAP: @KFILE @Stace_RNresists h...	2020-04-04 16:28:31	positive	google	1
1	RT @StarFeuri: I just drove to campus so that...	2020-04-04 16:28:31	positive	google	1
2	@marky_quinn I apologize I don't understand, ...	2020-04-04 16:28:31	negative	google	-1
3	RT @SpirosMargaris: Google's #AI teaches #rob...	2020-04-04 16:28:31	positive	google	1
4	@pablitasan Chocolate cake. Most recipes are ...	2020-04-04 16:28:30	positive	google	1
...	...	...	...	...	...
95	RT @GBCollege: ⚠ Attention Current Students! ...	2020-04-04 16:28:01	positive	google	1
96	RT @Trinhnomics: So Google has these mobility...	2020-04-04 16:28:01	positive	google	1
97	Do you have a question for Elliott Abrams, U....	2020-04-04 16:28:00	positive	google	1
98	RT @rasbt: "Swift: Google's bet on differenti...	2020-04-04 16:27:59	positive	google	1
99	@nothinghidden1 @ObamaMalik All of the patent...	2020-04-04 16:27:59	positive	google	1

Figure 4: Result of csv file generated

9 CONCLUSION

Using paper [1] as base paper it was possible to implement various machine learning algorithms to perform sentiment analysis of tweets . A comprehensive study of the comparison between the different models and their performance(accuracy) was also obtained. Reducing the dataset through feature extraction, enhanced the performance of the classifiers used and gave a better result.

References

[1] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-5.

[2] R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, 2016, pp. 64-67.

[3] Sumesh Kumar Nair, Ravindra Soni. "Sentiment Analysis On Movie Reviews Using Recurrent Neural Network" Iconic Research And Engineering Journals, 1(10)

[4] Zhang, L, Wang, S, Liu, B. Deep learning for sentiment analysis: A survey. WIREs Data Mining Knowl Discov. 2018; 8:e1253

[5] S. Rana and A. Singh, "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques," 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 2016, pp. 106-111.

[6] P. Jadon, D. Bhatia and D. K. Mishra, "A BigData approach for sentiment analysis of twitter data using Naive Bayes and SVM Algorithm," 2019 Sixteenth International Conference on Wireless and Optical Communication Networks (WOCN), Bhopal, India, 2019, pp. 1-6.

attach your review and visit log here.....

attach plagiarism report here.....