TWITTER SENTIMENT ANALYSIS

Seminar (IT290) Report

Submitted in partial fulfilment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

By

RAKSHATHA VASUDEV

(171IT131)

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY

KARNATAKA SURATHKAL, MANGALORE -575025

APRIL, 2019

D E C L A R A T I O N

I hereby *declare* that the *Seminar (IT290) Report* entitled TWITTER SENTIMENT ANALYSIS which is being submitted to the National Institute of Technology Karnataka Surathkal, in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in the department of Information Technology, is a ***bonafide report of*** the ***work carried out by me.*** The material contained in this project report has not been submitted to any University or Institution for the award of any degree.

RAKSHATHA VASUDEV
171IT131

(Name and Register Number of the Student)

Signature of the Student

Department of Information Technology

Place : NITK, SURATHKAL
Date : 03 / 04 / 2019

CERTIFICATE

This is to certify that the Seminar entitled "**TWITTER SENTIMENT ANALYSIS**" **has** been presented by Rakshatha Vasudev, a student of IV semester B.Tech. (I.T), Department of Information Technology, National Institute of Technology Karnataka, Surathkal, on 3rd April, during the even semester of the academic year 2018 – 2019, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology.

Examiner-1 Name

Signature of the Examiner-1 with Date

Examiner-2 Name

Signature of the Examiner-2 with Date

Guide Name

Signature of the Guide with Date

PLACE: NITK, Suratkal

DATE: 03/04/2019

Abstract

Social media have received more attention nowadays. Public and private opinion about a wide variety of subjects are expressed and spread continually via numerous social media. Twitter is one of the social media that is gaining popularity. Twitter offers organizations a fast and effective way to analyze viewers' perspectives towards the critical to success in the market place. Developing a program for sentiment analysis is an approach to be used to computationally measure peoples' perceptions. This paper reports on the design of a sentiment analysis, extracting a vast number of tweets. Experimental evaluations show that the proposed machine learning classifiers are efficient and perform better in terms of accuracy and time. The proposed algorithm is implemented in python and with the help of sci-kit learn. The results discusses the accuracy of the model built.

Table of Contents

List of Figures

Chapter 1: INTRODUCTION

Millions of people are using social network sites to express their emotions, opinion and disclose about their daily lives. However, people write anything such as social activities or any comment on products. Through the online communities provide an interactive forum where consumers inform and influence others. Moreover, social media provides an opportunity for business that giving a platform to connect with their customers such as social media to advertise or speak directly to customers for connecting with customer's perspective of products and services. In contrast, consumers have all the power when it comes to what consumers want to see and how consumers respond. With this, the company's success & failure is publicly shared and end up with word of mouth. However, the social network can change the behavior and decision making of consumers, for example, Figure 1.1 and 1.2 shows how important a users' review is to any kind of organization. So that, if organization can catch up faster on what their customer's think, it would be more beneficial to organize to react on time and come up with a good strategy to compete.

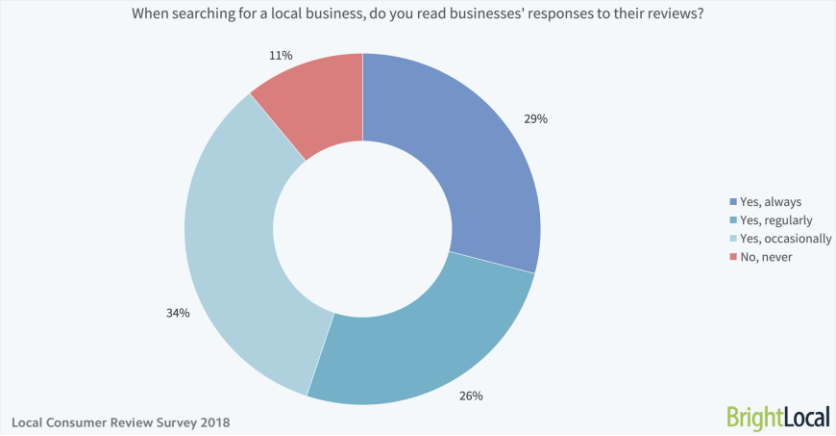

Figure 1.1: Survey on whether people look for the organization's response to reviews

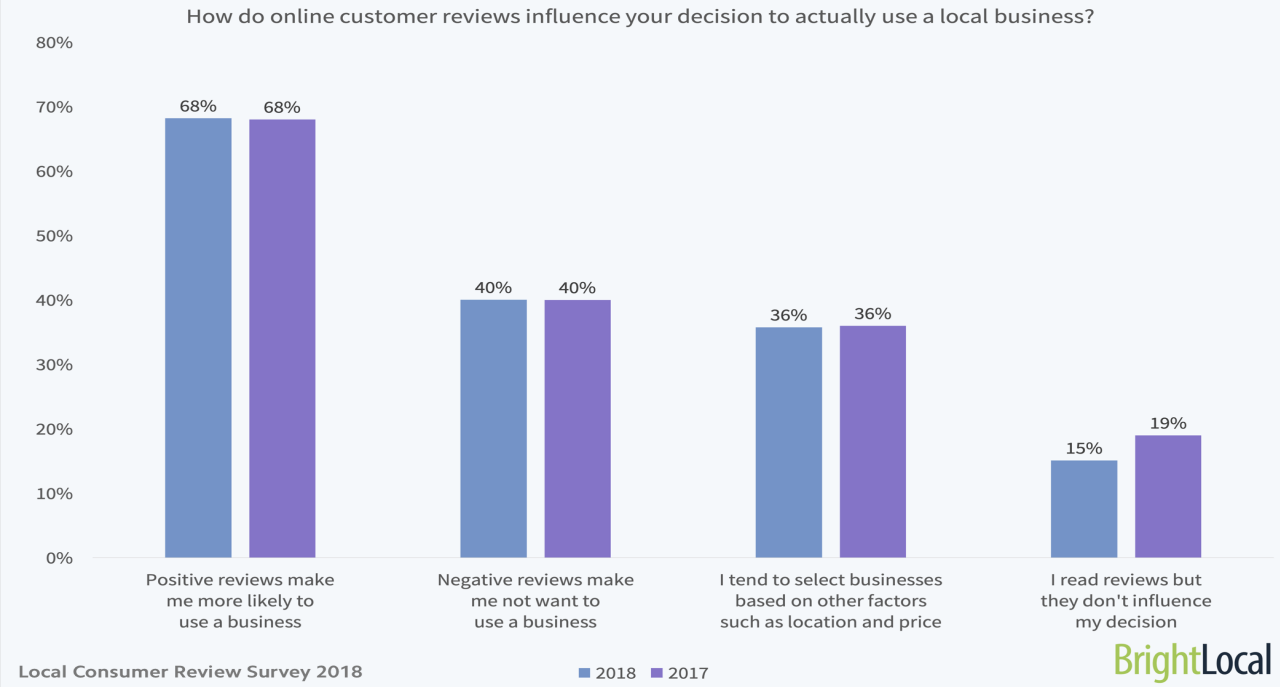

Figure 1.2: Survey on how much people are influenced by the reviews

Despite the availability of software to extract data regarding a person's sentiment on a specific product or service, organizations and other data workers still face issues regarding the data extraction.

-Sentiment Analysis of Web Based Applications Focus on Single Tweet Only. With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volumes of opinion texts in the form of tweets which is available for the sentiment analysis. This translates to a huge volume of information from a human viewpoint which make it difficult to extract a sentence, read them, analyze tweet by tweet, summarize them and organize them into an understandable format in a timely manner.

Emoticons, are a pictorial representation of human facial expressions, which in the absence of body language and prosody serve to draw a receiver's attention to the tenor or temper of a sender's nominal verbal communication, improving and changing its interpretation. For example, ☺ indicates a happy state of mind. Systems currently in place do not have sufficient data to allow them to draw feelings out of the emoticons. As humans often turn to emoticons to properly

express what they cannot put into words. Not being able to analyze this puts the organization at a loss. Short-form is widely used even with short message service (SMS). The usage of short-form will be used more frequently on Twitter so as to help to minimize the characters used.

1.1 Sentiment Analysis

Sentimental analysis is the process of computationally determining the opinion or attitude of the writers as positive, negative or neutral. Data mining is another name for sentimental analysis. In many fields like business, politics and public actions, determining the sentimental analysis is very important. Considering business, it is very useful to understand the customer ‘s feelings in order to develop their company. Next in politics: It can be even be used to predict the election results. There are two ways of classifications and they are (1) machine learning (2) lexicon-based approach. In this paper machine learning classifiers are implemented in sentimental analysis and is done in twitter because most of the politicians, famous personalities (even the president of various states) and even general people regularly update their moods in the form of tweets.

Chapter-2 Literature Review

2.2 Literature Review

In today's world, micro-blogging sites has become a platform for individuals or organizations across the world to express their opinions, sentiment and experience in the form of tweets, status updates, blog posts, etc. This platform has no political and economic restrictions. This paper discusses an approach where a dataset with random tweets are subjected to preprocessing and classified based on their emotional content as positive, negative and neutral. The performance of the algorithm is then analyzed. The paper concludes with the results (accuracy) of the model

2.1 Real life Application

Since the Opinion based or feedback-based application are more fashionable, now a days, the natural language processing community shows much interest in Sentiment Analysis and Opinion Mining system. The explosion of internet has changed the people 's life style, now they are more expressive on their views and opinions.

•Purchasing Product or Service: While purchasing a productor service, taking right decision is no longer a difficult task. By this technique, people can easily evaluate other 's opinion and experience about any product or service and also he can easily compare the competing brands. Now people don 't want to rely on external consultant. The Opinion mining and sentiment analysis extract people opinion form the huge collection of unstructured content, the internet, and analyze it and then present to them in highly structured and understandable manner

•Quality Improvement in Product or service: By Opinion mining and sentiment analysis the manufactures can collect the critic 's opinion as well as the favorable opinion about their product or service and thereby they can improve the quality of their product or service. They can make use of online product reviews from websites such as Amazon and C|Net, RottenTomatoes.com and IMDb.

•Marketing research: The result of sentiment analysis techniques can be utilized in marketing research. By sentiment analysis techniques, the recent trend of consumers about some product or services can be analyzed. Similarly, the recent attitude of general public towards some new government policy can also be easily analyzed. These all result can be contributed to collective intelligent research.

Chapter-3: TECHNICAL DISCUSSION

This chapter consists of four different sections. Section 3.1 clearly tells us about the proposed methodology that starts from the twitter data extractions and data pre-processing. The next section 3.2 tells about the feature extraction. The next section 3.3 tells about the machine learning algorithms. Section 3.4 has the result

Proposed Methodology

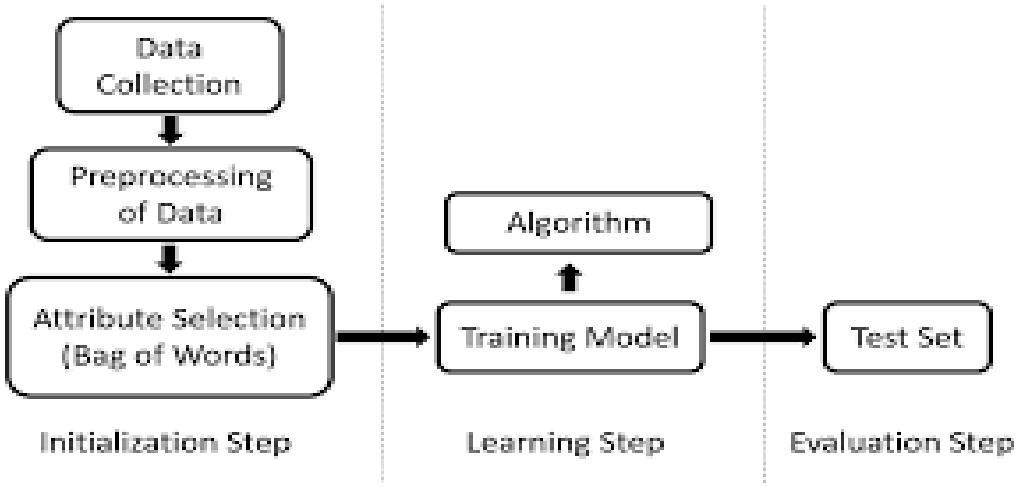

Figure 3.1: Proposed methodology

3.1 Data collection and Pre-processing

3.1.1 Dataset preparation

The dataset for the analysis was taken from SemEval '10(International workshop on Semantic Evaluation). It consists of tweets (random) and their corresponding classes (positive, neutral or negative). The rows which did not have tweets ('Not Available' tags) were removed.

A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas library can be used to create the data frame. The data which we use to train our model should not contain any noise. The dataset which we get are considered dirty(has unwanted words).In order to clean it ,text pre-processing is done

3.1.2 Using nltk

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion-forum.

- Removing Stop words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words taking up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that is consider to be stop words using nltk. (stopwords) where stopwords is the list.
- Stemming: Stemming is the process of producing morphological variants of a root/base word. A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”. This is done by Porterstemmer () class.
- Lemmatization: Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So, it links words with similar meaning to one word.

3.1.3 Using string. Punctuation

A string contains letters, whitespace, numbers. And it has punctuation: these characters include commas and periods and semicolons. With Python**,** we can access the string. punctuation constant. This contains all the common punctuation characters.

- Removing punctuation
- Removing user handles (@)

3.1.4 Label Encoding

Encode labels with value between 0 and n_classes-1 using sklearn library. This is done to convert categorical data, or text data, into numbers, which our predictive models can better understand. This is done by sklearn's LabelEncoder.

3.2 Feature Extraction

Selection of useful words from the tweet is called as feature extraction. In the feature extraction method, we extract the aspects from the pre-processed twitter dataset. Correct feature selection techniques are used in sentiment analysis that has got a significant role for identifying relevant attributes and increasing classification (machine learning) accuracy

3.2.1 Bag of Words (BoW)

Bag of Words (BOW) is a method to extract features from text documents. These features can be used for training machine learning algorithms. It creates a vocabulary of all the unique words occurring in all the documents in the training set. In simple terms, it's a collection of words to represent a sentence with word count and mostly disregarding the order in which they appear.

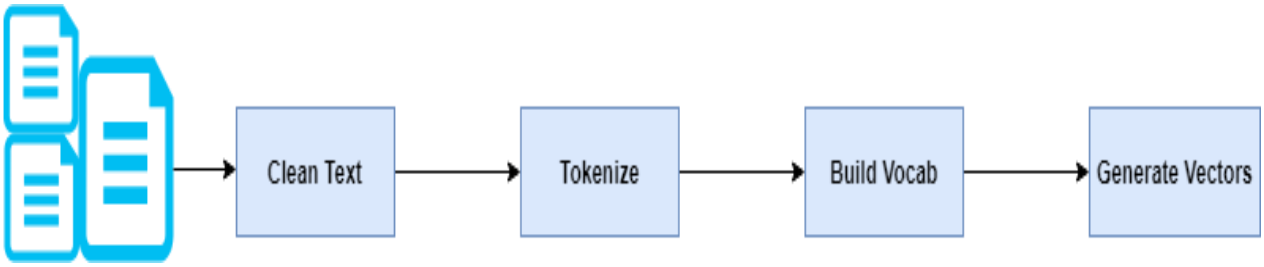

Figure 3.2: Bag of Words to generate word vectors

The pre-processed data is divided using train_test_split into training and test corpus. (X_train, X_test, y_train and y_test). The feature is extracted from the training corpus.

The procedure involves Tokenization, counting, normalization. The strings are split into tokens and given unique id. Each of their frequency is calculated and the weights are normalized as some of the words are discarded. Generated vectors can be input to your machine learning algorithm.

3.3 Classification

Machine learning is the study of algorithms that can learn from and make predictions on data. It is also called as related to prediction-making on some data. There are many machine learning

algorithms. They are used to classify the tweets/texts to their corresponding classes. This paper discusses one of the simplest algorithms -KNN Classifier

3.3.1 K-Nearest Neighbors

It is commonly used for its easy of interpretation and low calculation time.

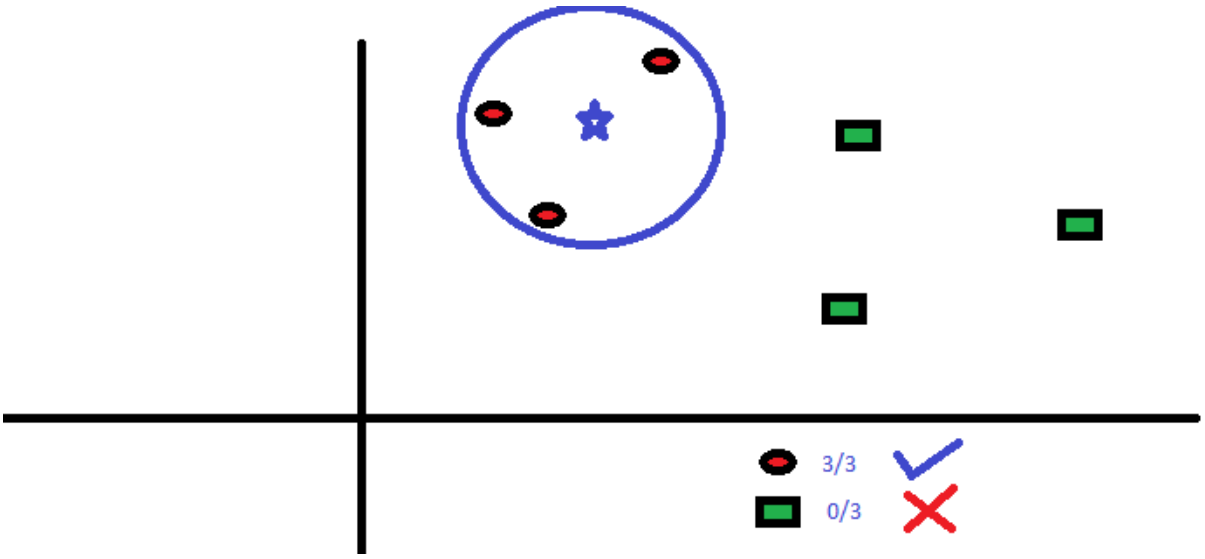

Figure 3.3: KNN visualization

The figure above is a spread of red circles (RC) and green squares (GS). We intend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else. The "K" is KNN algorithm is the nearest neighbors we wish to take vote from. Let's say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three datapoints on the plane. The three closest points to BS are all RC. Hence, with good confidence level we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm.

X_train and y_train are used to fit the model. The model is predicted by using the X_test and is compared with y_test for the accuracy.

3.3.2 **Determining K:** So, the value of K seems to be very crucial.

```
: Text(0, 0.5, 'Mean Error')
```

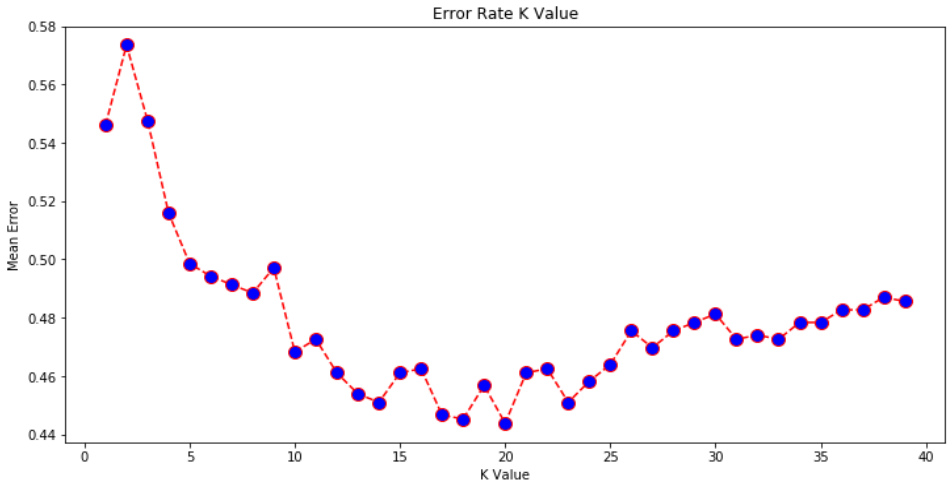

Figure 3.4: K decision

The K value for this model has been decided by plotting mean error v/s a range of K values graph.

3.4 **Result**:

The model resulted in an accuracy of 55.61% with a score of 58.93%.

```
[20]: from sklearn.neighbors import KNeighborsClassifier
      model=KNeighborsClassifier(n_neighbors=20)
      #print(train.shape,trainlabel.shape)
      model.fit(bowTrain,y_train)
      x=model.predict(bowTest)
      print(model.score(bowTrain,y_train))

      0.5893944248638257

[21]: count=0
      for i in range(len(y_test)):
          if y_test[i]==x[i]:
              count=count+1
      print('accuracy:',count/len(y_test)*100)

      accuracy: 55.61959654178674
```

Figure 3.5: Result

Chapter-4: Conclusion and Future Scope

Sentiment analysis is an emerging field in decision making process and is developing fast. The future development will be Less to do with improving the accuracy of the algorithms, but instead focus on the area of determining where you can correlate sentiment with behavior

There is a huge scope of improvement of these existing sentiment analysis models. Existing sentiment analysis models can be improved further with more semantic and commonsense knowledge.

4.2 References

[1] S. Siddharth, R. Darsini, Dr. M. Sujithra (2018),” Sentiment Analysis on Twitter Data Using Machine Learning Algorithms in Python”, *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 5, Issue 2*

[2] Avinash Navlani (2017) -” KNN Classification using Scikit-learn “, https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn (28/3/2019)

[3] Jason Brownlee (2017)-” A Gentle Introduction to the Bag-of-Words Model”, https://machinelearningmastery.com/gentle-introduction-bag-words-model/ (5/3/2019)

[4] Praveen Dubey (2018)-” An introduction to B*ag of Words and how to code it* in Python for NLP”, https://medium.freecodecamp.org/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04 (5/3/2019)

[5] Tavish Shrivastava (2014)-” Introduction to k-Nearest Neighbors: Simplified (with implementation in Python)”, https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/ (27/3/2019)