Twitter Sentiment Analysis

B Deva Sai Hritvik : 11802864

B Mounika : 11814710

Y V Sriram : 11803046

V V Ganesh : 11802955

T Sai Santhosh Vardhan : 11802338

ABSTRACT

A place where most of our opinions are supressed, where people couldn’t voice out their opinion, where they couldn’t be vocal, where they haven’t been heard is what upsets us. With the gaining popularity of the social media, people are able to speak what’s bothering them, can connect with people and reconnect with themselves. One such platform is Twitter. People can literally tweet about anything and everything. Any news, any incident, any update is on twitter once its official. With gaining popularity its been on of the most widely used platform with a user base of over 300 million. With this huge load of data we’ve formed the idea of sentiment analysis. This project deals with all the unstructured data on twitter and classify them where the opinions are positive , negative or neutral based on the type of data and the content preferred, into meaningful representable format. In this paper we provide and devise a method to classify this data using Naïve Bayes and R programming and also discuss the setbacks and application of Twitter sentiment analysis.

1.INTRODUCTION

Today, the age of the Internet has changed the way people work. Express their opinions and opinions. It's mostly passing now .Blog posts, online forums, product review sites, social media Media and more Today, millions of people are using social media. Popular social media websites like Instagram , Twitter and Facebook. Express their feelings and opinions and share their opinions in everyday life. Through the online community we get one interactive media where consumers inform and influence others through the forum. Social media is a lot make a wealth of data emotional in the form of tweets, status updates, and blogs Posts, comments, reviews, etc. Further social media Is providing an opportunity for enterprises by providing a platform contacting their customers for advertising. Mostly people dependent on user-generated content online Scope of decision making. For instance if anyone wants to buy, If someone wants to use a product or service, they look for it's reviews online, discuss on social media and friends before taking decision. Too much user-generated content analyse regular users. Therefore, it needs to be automated

2. Sentiment Analysis

Sentiment analysis is a technique used in text mining. Twitter sentiment analysis means using advanced text mining technology to analyse the mood of text (here, tweets) in positive, negative, and neutral formats. Also known as opinion mining, its main purpose is to analyse conversations, opinions and exchanges (all in the form of tweets) to determine business strategy, political analysis and assess public behaviour. R and Python are commonly used for Twitter datasets for sentiment analysis. Sentiment analysis is a technique used in text mining. Therefore, it can be used as a text mining technique to analyse the underlying emotions of text messages. H. Of tweets. Twitter's feelings and opinions can be positive, negative, or neutral. However, no algorithm can provide 100% accuracy or prediction when analysing emotions.

Sentiment analysis is the identification and classification of opinions and emotions expressed in the original text. Social media produces large amounts of emotional data in the form of tweets, status updates, blog posts, and more.

Sentiment evaluation of this consumer generated facts may be very beneficial in understanding the opinion of the crowd. Twitter sentiment evaluation is tough in comparison to preferred sentiment evaluation because of the presence of slang phrases and misspellings. The most restrict of characters which can be allowed in Twitter is 140. Knowledge base technique and Machine gaining knowledge of technique are the 2 techniques used for studying sentiments from the text. In this paper, we attempt to research the twitter posts approximately digital merchandise like mobiles, laptops and many others the usage of Machine Learning technique. By doing sentiment evaluation in a particular area, it's miles viable to become aware of the impact of area statistics in sentiment classification. We gift a brand new function vector for classifying the tweets as positive, poor and extract peoples` opinion approximately merchandise. Textual Information retrieval strategies in particular consciousness on processing, looking or reading the real statistics present. Facts have an goal element but, there are a few other text contents which explicit subjective characteristics. This content is in particular opinions, sentiments, appraisals, attitudes, and emotions, which shape the centre of Sentiment Analysis(SA). It gives many hard possibilities to increase new applications, in particular because of the massive increase of available statistics on on-line reasserts like blogs and social networks. For instance, hints of objects proposed with the aid of using a device may be expected with the aid of using taking into account concerns including wonderful or terrible opinions approximately the ones objects with the aid of using utilising SA.

2.1 Objective

The main objective of the application is to extract data from the twitter and clean it , then analysis and visualize the result. Finally giving the overall result whether the post got positive or negative reviews.

2.2 Data Pre-processing

The tweet contains many opinions about the data there it is represented in different ways by different users. Twitter dataset used in this study has already been split in two classes or negative and positive polars, therefore sentiment analysis of data is easier to observe effect of various functions.

Tweets are imported using R and the data is cleaned by removing emoticons and URLs. Lexical Analysis as well as Naive Bayes Classifier is used to predict the sentiment of tweets and subsequently express the opinion graphically through

gg plots, histogram, pie chart, word cloud and tables.

1. We remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic)

2. We check the spellings

3. We replace all the emoticons with their

sentiment.

4. We remove all punctuations, symbols, and numbers

2.3 Feature Extraction

The applications for sentiment analysis are endless. It is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics However, it is also practical for use in business analytics and situations in which text needs to be analyzed. Preprocessed datasets have many characteristics. In the feature extraction method extracts aspects from the processed record. Later on using this aspect, positive and negative polarities of useful sentences determine the opinions of people with the following models graphs and plots.

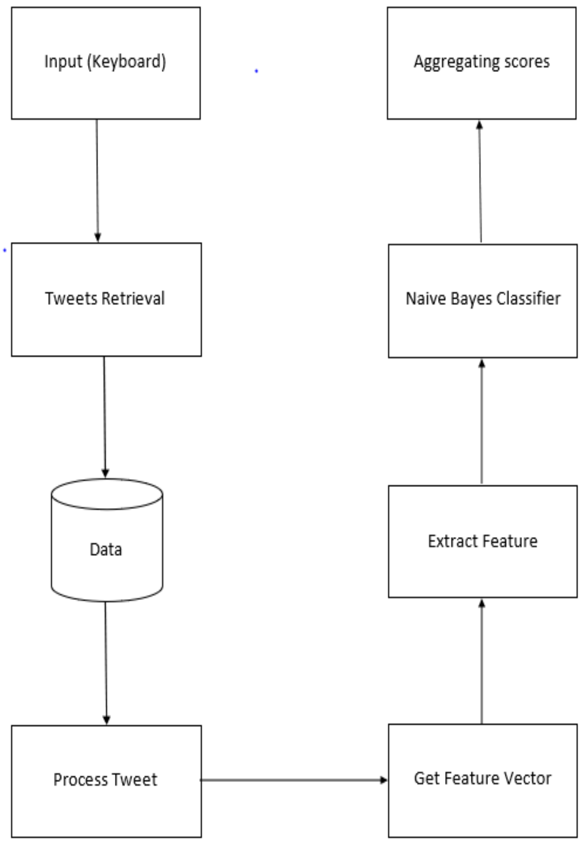
2.4 Training

Supervised learning is an important technique for solving classification problems. Training the classifier makes it easier for future predictions for unknown data.

3. Classification

Naïve Bayes

The Naïve Bayes classifier is the simplest and most commonly used classifier. Naïve Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a label.



P(*label*) is the prior probability of a label or the likelihood that a random feature set the label. *P*(*features*|*label*) is the prior probability that a given feature set is being classified as a label. *P*(*features*) is the prior probability that a given feature set is occurred. Given the Naïve assumption which states that all feature sare independent, the equation could be rewritten as follows:

$$P(label|features) = \frac{P(label) * P(features\ |label)}{P(features)}$$

Algorithm Dictionary: Generation:

Count occurrence of all word in our whole data set and make a dictionary of some most frequent words.

Feature Set Generation

All document is represented as a feature vector over the space of dictionary words. For each document, keep track of dictionary words along with their number of occurrences in that document.

4. Implementation

There are primarily two types of implementations for sentiment classification of opinionated texts:

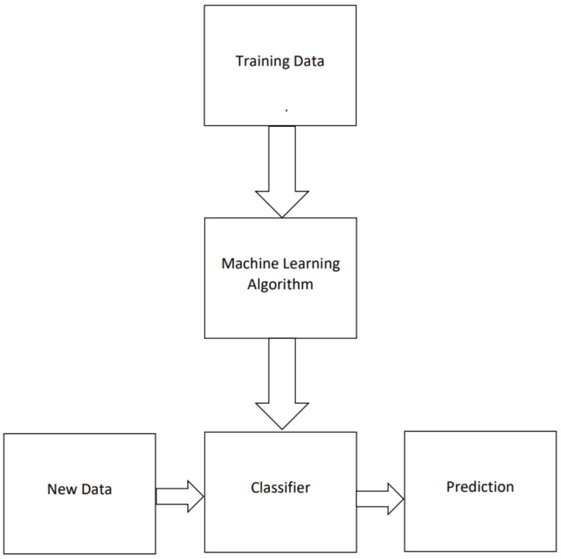
Using a Machine learning based text classifier such as Naive Bayes

Using Natural Language Processing

We will be using those machine learning and natural language processing for sentiment analysis of tweet.

Machine Learning

It is based on text classifiers, they are a kind of superintended machine learning model, where the classifier demands to be guided by some labelled training data be for it can be applied to actual classification task. The training data is usually an extracted portion of the original data hand labelled manually. After suitable training they can be used on the actual test data. The Naive Bayes is a statistical classifier whereas Support Vector Machine is a kind of vector space classifier. The statistical text classifier scheme of Naïve Bayes (NB)can be adapted to be used for sentiment classification problem.



Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computer and human(natural)languages. This approach utilizes the publicly available library of Senti Word Net, which provides a sentiment polarity values for every term occurring in the document. In this lexical resource each term t occurring in WordNet is associated to three numerical scores obj(t), pos(t)and neg(t), describing the objective, positive and negative polarities of the term, respectively. These three scores are computed by combining the results produced by eight ternary classifiers. Word Netisalar gelexical analysis of the database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing. Its groups words together based on their meanings. Synet is nothing but a set of one or more Synonyms. This approach uses Semantics to understand the language.

Major tasks in NLP that helps in extracting sentiment from a sentence:

1) Extracting part of the sentence that reflects the sentiment
2) Understanding the structure of the sentence
3) Different tools which help process the textual data

Basically, Positive and Negatives cores got from Senti Word Net according to its part-of-speech tag and then by counting the total positive and negative scores we determine the sentiment polarity based on which class (i.e. either positive or negative) has received the highest score.

As it can be visualize data to class text classification problem: in positive and negative classes. Support Vector machine (SVM) is a kind of vector space model-based classifier which requires that the text documents should be transformed to feature vectors before they are used for classification. Usually, the text documents are transformed to multidimensional vectors. The entire problem of classification is then classifying every text document represented as a vector into a class. It is at a type of large margin classifier. Here the goal is to find a decision boundary between two classes that is maximally far from any document in the training data.

This approach needs

i) A good classifier such as Naïve Byes
ii) A training set for each class

There are various training sets available on Internet such as Movie Reviews data set, twitter dataset, etc. Class can be Positive, negative. For both the classes we need training data sets.

4. Evaluation

The performance of sentiment classification can be evaluated by using four indexes calculated as the following equations:

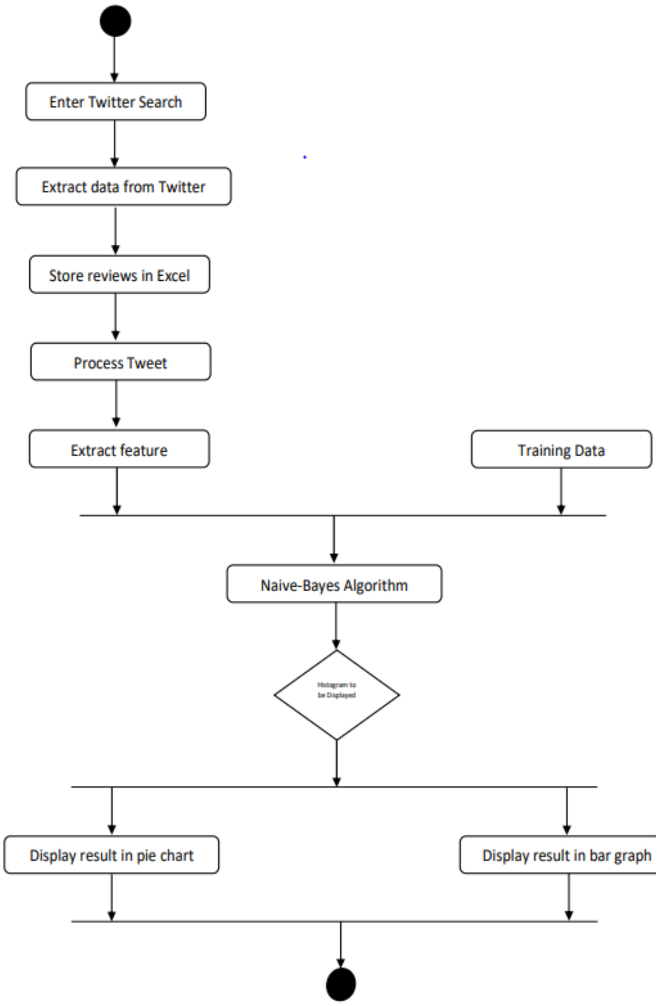
Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP)

Recall = TP/(TP+FN) F1 = (2×Precision×Recall)/(Precision+Recall)

In which TP, FN, FP and TN refer respectively to the number of true positive instances, the number of false negative instances, the number of false positive instances and the number of true negative instances.

	Predicted Positives	Predicted Negatives
Actual Positive.	TP	FN
Actual Negative	FP	TN


Enter Twitter Search
Extract data from Twitter
Store reviews in Excel
Process Tweet
Extract feature
Training Data
Naive-Bayes Algorithm
Histogram to be Displayed
Display result in pie chart
Display result in bar graph