A PROJECT REPORT

ON

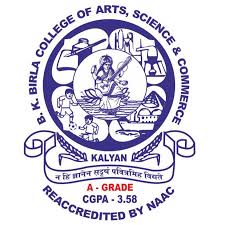
“**TWITTER DATA SENTIMENT ANALYSIS**”

SUBMITTED BY

**Mr. Pinku Phulchand Varma**

UNDER THE GUIDANCE OF

**PROF. MRS. KAJAL JAISINGHANANI**



**DEPARTMENT OF COMPUTER SCIENCE**

**B. K. BIRLA COLLEGE OF ARTS, SCIENCE &**

**COMMERCE (AUTONOMOUS), KALYAN**

**AFFILIATED TO UNIVERSITY OF MUMBAI**

**(2020 - 21)**

**TWITTER DATA SENTIMENT ANALYSIS**

|  |  |  |
| --- | --- | --- |
| **Sr.No** | **Content** | **Page No.** |
| **1** | **ABSTRACT** |  |
| **2** | **INTRODUCTION** |  |
| **3** | **MOTIVATION** |  |
| **4** | **APPLICATION** |  |
| **5** | **REQUIRMENTS SPECIFICATION** |  |
| **6** | **RELATED WORK** |  |
| **7** | **OBJECTIVES** |  |
| **8** | **METHODOLOGY** |  |
| **9** | **CONCLUSION** |  |
|  |  |  |

**ABSTRACT**

Twitter is an online microblogging and social networking platform, which allows users to write short status, updates of maximum length 280 characters. These tweets reflect public sentiment about various topics and events happening.

Analysing the public sentiment can help, firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. Sentiment analysis techniques are widely

popular for this purpose. In this paper, we have tried to define and compare various sentiment classification approaches/methods for finding out the sentiments behind the

tweet.

**INTRODUCTION :-**

Opinion and sentimental mining is an important research areas because due to the huge number of daily posts on social networks, extracting people’sopinion is a challenging task. About 90 percent of today’s data has been provided during the last two years and getting insight into this large-scale data is not trivial.

Sentimental analysis has many applications for different domains for example in businesses to get feedbacks for products by which companies can learn user's feedback and reviews on social medias.

Opinion and sentimental mining have been well studied in this reference and all different approaches and research fields have been discussed. There are also some works have been done on Facebook sentimental analysis however in this paper we mostly focus on the Twitter sentimental analysis.

For a larger texts one solution could be understand the text, summarize it and give weight to it whether it is positive, negative or neutral. Two fundamental approaches to extract text summarization are an extractive and abstractive method. In the extractive method, words and word phrases are extracted from the original text to generate a summary. In an abstractive method, tries to learn an internal language representation and then generates summary that is more similar to the summary done by human.

**PROBLEM STATEMENT:-**

• The problem in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level .

• whether the expressed opinion in a document, asentence or an entity feature/aspect is positive, negative, or neutral.

**MOTIVATION :-**

• An aspect of social media data such as Twitter messages is that it includes rich structured information about the individuals involved in the communication.

• It can lead to more accurate tools for extracting semantic information.

• It provides means for empirically studying properties of social interactions.

• Freely available, annotated corpus, Pre-written Classifier Codes in Python using Jupyter.

**Applications -**

Sentiment Analysis Dataset Twitter has a number of applications:

**Business:** Companies use Twitter Sentiment Analysis to develop their business strategies, to assess customers’ feelings towards products or brand, how people respond to their campaigns or product launches and also why consumers are not buying certain products.

**Politics:** In politics Sentiment Analysis Dataset Twitter is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Sentiment Analysis Dataset Twitter is also used for analyzing election results.

**Public Actions:** Twitter Sentiment Analysis also is used for monitoring and analyzing social phenomena, for predicting potentially dangerous situations and determining the general mood of the blogosphere.

**Software** **Requirements -**

**Operating System:-**windows 7 ,windows XP / vista ,

windows 8 and higher versions

**Language** :-Anacondas Jupyter Notebook(python)

**Hardware Requirement-**

**RAM:**- 1GB and more

**Processor :**- Any intel processor

**Hard Disk :**- 6GB and more

**Speed :**- 1GHZ and more

**RELATED WORK:**

The related work section may also be called a literature review. The point of the section is to highlight work done by others that somehow ties in with your own work. It may be work that you’re basing your work off of, or work that shows others attempts to solve the same problem.

**1.Twitter as a Corpus for Sentiment Analysis and Opinion Mining.**

Microblogging today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life everyday. Therefore microblogging web-sites are rich sources of data for opinion mining and sentiment analysis. Becausemmicroblogging has appeared relatively recently, there are a few research works that were devoted to this topic. In our paper, we focus on using Twitter, the most popular microblogging platform, for the task of sentiment analysis. We show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. We perform linguistic analysis of the collected corpus and explain discovered phenomena.

Using the corpus, we build a sentiment classifier, that is able to determine positive, negative and neutral sentiments for a document. Experimental evaluations show that our proposed techniques are efficient and performs better than previously proposed methods. In our research, we worked with English, however, the proposed technique can be used with any other language.

In our research, we have presented a method for an automatic collection of a corpus that can be used to train a sentiment classifier. We used TreeTagger for POS-tagging and observed the difference in distributions among positive, negative and neutral sets. From the observations we conclude that authors use syntactic structures to describe emotions or state facts. Some POS-tags may be strong indicators of emotional text.

We used the collected corpus to train a sentiment classifier. Our classifier is able to determine positive, negative and neutral sentiments of documents. The classifier is based onthe multinomial Na¨ıve Bayes classifier that uses N-gram and POS-tags as features.

As the future work, we plan to collect a multilingual corpus of Twitter data and compare the characteristics of the corpus across different languages. We plan to use the obtained data to build a multilingual sentiment classifier.

**2.Language-Independent Twitter Sentiment Analysis**

Sentiment classification can benefit companies by providing data for analyzing customer feedback for products or conducting market research.

Sentiment classifiers need to be able to handle tweets in multiple languages to cover a larger portion of the available tweets. Traditional classifiers are however often language specific and require much work to be applied to a different language. We analyze the characterstics and feasibility of a language-independent, semisupervised sentiment classification approach fortweets.

We use emoticons as noisy labels to generate training data from a completely raw set of tweets. We train a Na¨ıve Bayes classifier on our data and evaluate it on over 10000 tweets in 4 languages that were human annotated using the Mechanical Turk platform. As part of our contribution, we make the sentiment evaluation dataset publicly available.In this paper we presented a language-independent classification approach to detect the sentiment polarity in tweets.

Our approach uses a semi-supervised heuristic labeling scheme to acquire large amounts training data in a variety of languages, and content-based features that work well across languages. In our experiments, we trained a Na¨ıve Bayes classifier on source sets of millions of tweets in English, German, French and Portuguese. We compared the results between languages using tweets that were humanannotated using the Amazon Mechanical Turk service. We make this multilingual evaluation dataset publicly available.

We showed that the used algorithm achieves a good performance for tweets of multiple languages, especially given its efficient applicability to new languages. The classifier can be trained effortlessly on new languages, given only raw training data. We have not dealt with classifying subjectivity in tweets in this work. This is something we hope to be able to accomplish in future work.

**3. A Study on Sentiment Analysis Techniques of Twitter Data.**

The entire world is transforming quickly under the present innovations. The Internet has become a basic requirement for everybody with the Web being utilized in every field. With the rapid increase in social network applications, people are using these platforms to voice them their opinions with regard to daily issues. Gathering and analyzing peoples’ reactions toward buying a product, public services, and so on are vital. Sentiment analysis (or opinion mining) is a common dialogue preparing task that aims to discover the sentiments behind opinions in texts on varying subjects. In recent years, researchers in the field of sentiment analysis have been concerned with analyzing opinions on different topics such as movies, commercial products, and daily societal issues. Twitter is an enormously popular microblog on which clients may voice their opinions. Opinion investigation of Twitter data is a field that has been given much attention over the last decade and involves dissecting “tweets” (comments) and the content of these expressions.

As such, this paper explores the various sentiment analysis applied to Twitter data and their outcomes.Interesting area for future study includes the fluctuations in the performance of sentiment analysis algorithms in cases where multiple features are considered. In other words, combining various features was found to lead to improve the performance in most cases, but substandard performance in others. Thus, an exploration into the causes of these performance instabilities would be an intriguing direction for future works. Another might be to investigate the data sparsity issue using both ensemble and hybrid approaches. The intention behind this is to measure the robustness of various Twitter sentiment approaches the data sparsity. A further area of study might be the utilization of active learning techniques to detect Twitter sentiments and to increase the confidence of decision makers.

**4.Sentiment Analysis of Twitter Data: A Survey of Techniques**

This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. In this paper, we provide a survey and a comparative analyses of existing techniques for opinion mining like machine learning and lexicon-based approaches, together with evaluation metrics. Using various machine learning algorithms like Naive Bayes, Max Entropy, and Support Vector Machine, we provide research on twitter data streams.We have also discussed general challenges and applications of Sentiment Analysis on Twitter.

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction. The words opinion, sentiment, view and belief are used interchangeably but there are differences between them.

* **Opinion:** A conclusion open to dispute (because different experts have different opinions )
* **View:** subjective opinion
* **Belief:** deliberate acceptance and intellectual assent
* **Sentiment:** opinion representing one„s feelings

In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with cross domain and cross-lingual methods and some evaluation metrics. Research results show that machine learning methods, such as SVM and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few effort in human-labeled document .

We also studied the effects of various features on classifier. We can conclude that more the cleaner data, more accurate results can be obtained. Use of bigram model provides better sentiment accuracy as compared to other models. We can focus on the study of combining machine learning method into opinion lexicon method in order to improve the accuracy of sentiment classification and adaptive capacity to variety of domains and different languages.

**5.Sentimentor: Sentiment Analysis of Twitter Data**

Social networks have revolutionised the way in which people communicate. Information available from social networks is beneficial for analysis of user opinion, for example measuring the feedback on a recently released product, looking at the response to policy change or the enjoyment of an ongoing event. Manually sifting through this data is tedious and potentially expensive. Sentiment analysis is a relatively new area, which deals with extracting user opinion automatically. An example of a positive sentiment is, “natural language processing is fun” alternatively, a negative sentiment is “it’s a horrible day, i am not going outside”. Objective texts are deemed not to be expressing any sentiment, such as news headlines, for example “company shelves wind sector plans”.

There are many ways in which social network data can be leveraged to give a better understanding of user opinion such problems are at the heart of natural language processing (NLP) and data mining research. In this paper we present a tool for sentiment analysis which is able to analyse Twitter data. We show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. Using the corpus we build a sentiment classifier, that is able to determine positive, negative and objective sentiments for a document. In this paper we have presented a way in machine learning techniques can be applied to large sets of data to establish membership, in this case positivity, negativity and objectivity. We have looked at common process in NLP that can help us derive the meaning or context of a given phrase. We have demonstrated how to collect an original corpus for sentiment classification and the refinement that is needed with such data.

**Objectives:**

Sentiment analysis(also known as opinion mining) refers to the use of natural language processing,text analysis and computational linguistics to identify and extract subjective information in source materials.Consumers can use sentiment analysis to research product and service s before a purchase.Production companies can use the public opinion to determine acceptance of their products and the public demand.

Movie-goers can decide whether to watch a movie or not after going through other people’s reviews.

Twitter sentiment analysis allows you to listen to your customers and understand what they need. By introducing sentiment analysis tools into your workflows, you can automatically organize [unstructured information](https://monkeylearn.com/unstructured-data/) (which includes Twitter data) in real-time, at scale, and accurately:

* **Scalability:** Analyze hundreds or thousands of tweets mentioning your\
* brand and automate manual tasks. Easily scale sentiment analysis tools as your data grows and gain valuable insights on the go.
* **Real-Time Analysis:** Twitter sentiment analysis is essential for monitoring sudden shifts in customer moods, detecting if complaints are on the rise, and for taking action before problems escalate. With sentiment analysis, you can monitor brand mentions on Twitter in real-time and gain actionable insights.
* **Consistent Criteria:** Avoid inconsistencies that stem from several agents tagging data against different criteria. Instead, train a machine learning model to perform sentiment analysis, using one set of rules, on all your Twitter data, so results are consistent.

**METHODOLOGY:**

**PROPOSED SYSTEM :-**

In this system, we are using machine learning based approach for sentiment classification. For this, we are constructing dataset of tweets, which are obtained from Twitter using Tweepy API. After obtaining tweets, they are pre-processed to remove the noise. The tweets are labelled as either positive, negative or neutral. After pre-processing, useful and significant features are extracted from tweets. The machine learning classifiers are applied on the training dataset. The model obtained from training, is applied on unseen test dataset to check the accuracy of the model. A web application will be created which will display the results of the classification. The

results are visualized and displayed on website for user convenience.

**Data Collection**

Twitter allows researchers to collect tweets by using a Twitter API. To collect tweets from twitter one must have a twitter account to obtain twitter credentials (i.e. API key, API secret, Access token and Access token secret) which can be obtained from twitter developer site. Then the user needs to install a library to connect to the twitter API using these credentials. Now tweets can be extracted from twitter.

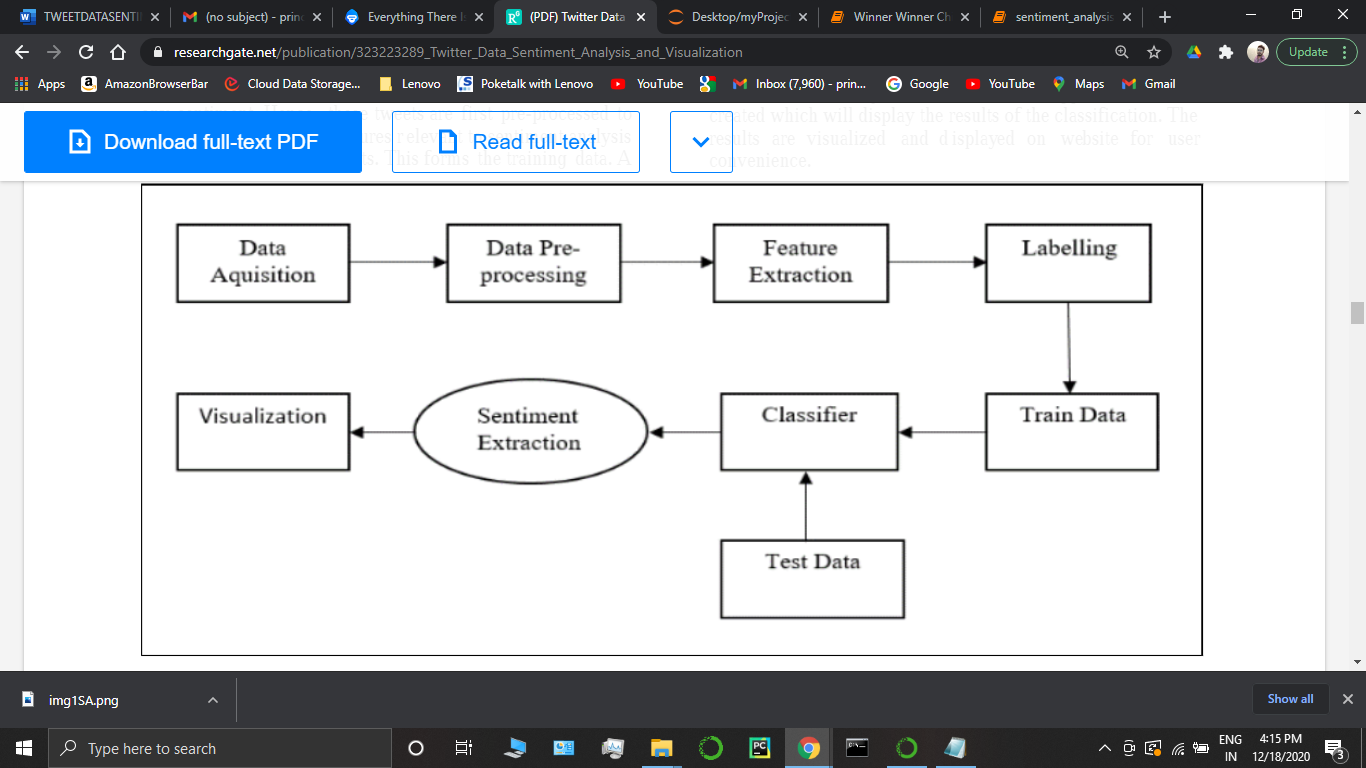


Fig: System Architecture

## **Pre-processing Tweets**

This is one of the essential steps in any natural language processing (NLP) task. Data scientists never get filtered, ready-to-use data. To make it workable, there is a lot of processing that needs to happen.

* **Letter casing:** Converting all letters to either upper case or lower case.
* **Tokenizing:** Turning the tweets into tokens. Tokens are words separated by spaces in a text.
* **Noise removal:** Eliminating unwanted characters, such as HTML tags, punctuation marks, special characters, white spaces etc.
* **Stopword removal:** Some words do not contribute much to the machine learning model, so it's good to remove them. A list of stopwords can be defined by the nltk library, or it can be business-specific.
* **Normalization:** Normalization generally refers to a series of related tasks meant to put all text on the same level. Converting text to lower case, removing special characters, and removing stopwords will remove basic inconsistencies. Normalization improves text matching.
* **Stemming:** Eliminating affixes (circumfixes, suffixes, prefixes, infixes) from a word in order to obtain a word stem. Porter Stemmer is the most widely used technique because it is very fast. Generally, stemming chops off end of the word, and mostly it works fine.
  + *Example: Working -> Work*
* **Lemmatization:** The goal is same as with stemming, but stemming a word sometimes loses the actual meaning of the word. Lemmatization usually refers to doing things properly using vocabulary and morphological analysis of words. It returns the base or dictionary form of a word, also known as the *lemma* .
  + *Example: Better -> Good.*
* **Vectorizing Data:** Vectorizing is the process to convert tokens to numbers. It is an important step because the machine learning algorithm works with numbers and not text.In this guide, you'll implement vectorization using tf-idf. There are other techniques as well, such as Bag of Words and N-grams.

## **Story Generation and Visualization from Tweets**

In this section, we will explore the cleaned tweets text. Exploring and visualizing data, no matter whether its text or any other data, is an essential step in gaining insights. Do not limit yourself to only these methods told in this tutorial, feel free to explore the data as much as possible.

Before we begin exploration, we must think and ask questions related to the data in hand. A few probable questions are as follows:

* What are the most common words in the entire dataset?
* What are the most common words in the dataset for negative and positive tweets, respectively?
* How many hashtags are there in a tweet?
* Which trends are associated with my dataset?
* Which trends are associated with either of the sentiments? Are they compatible with the sentiments?

### **Understanding the common words used in the tweets: WordCloud**

Now I want to see how well the given sentiments are distributed across the train dataset. One way to accomplish this task is by understanding the common words by plotting wordclouds.

*A wordcloud is a visualization wherein the most frequent words appear in large size and the less frequent words appear in smaller sizes.*

### **Understanding the impact of Hashtags on tweets sentiment**

Hashtags in twitter are synonymous with the ongoing trends on twitter at any particular point in time. We should try to check whether these hashtags add any value to our sentiment analysis task, i.e., they help in distinguishing tweets into the different sentiments.

For instance, given below is a tweet from our dataset:



**Machine Learning Classifier:**

Once the features are extracted and training dataset is formed, next comes the step in which a machine learning classifier is applied on the training dataset. There are various supervised as well as unsupervised machine-learning classifiers for sentiment analysis. The model obtained from the training dataset is applied on the unseen test dataset, to check the accuracy and performance of the model. Supervised classifiers such as Naïve Bayes classifier, Support Vector Machine, decision tree algorithms can be used for classification of sentiments.

**Naïve Bayes Classifier:**

Naive Bayes classifier works very well for text classification as it computes the posterior probability of a class, based on the distribution of the words (features) in the document. The model uses the Bag of words feature extraction. Naïve Bayes classifier assumes that each feature is independent of each other. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

P(label | features ) = P(features | label) \* P(label) / P(features)

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.

**Support Vector Machine**

The main principle of SVMs is to determine linear separators in the search space, which can best separate the different classes. There can be several hyperplanes, but SVM classifier chooses the one, which gives the maximum distance for any point. The hyperplane chosen should depict maximum margin of separation. Text classification are perfectly suited for SVMs because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories.

**CONCLUSION AND FUTURE SCOPE**

This study proposes a sentiment analysis system using machine-learning approach. The study is based on sentence level sentiment classification. It is useful for finding out sentiments of people regarding any topic or event. It is also useful in predicting socio-economic phenomena like stock market prediction. As a future work, this study can be expanded to include feature/aspect level classification, which is useful in product review and recommendation system. The number of sentiment classes can be increased to get more refined sentiment prediction. The study is a prototype and is meant to present the potential use of social networking platforms such as Twitter for large scale information gathering and processing for future social media related applications. The study can be extended for applications such as emergency management, social unrest etc. The study can be extended.