**Chapter 1**

**Introduction**

In the last couple of years the social medium Twitter has become more and more popular, since Twitter is the most used micro blogging website with about 500 million active users and 340 million tweets a day; it is an interesting source of information. The messages, or in Twitter terms the tweets, are a way to share interests publicly or among a denned group. Twitter distinguishes itself from other social media by the limited message size. The maximum size of 140 characters restricts users in their writing. Twitter is therefore challenging their users to express their view in one or two key sentences.

Because Twitter is widely adopted through all strata, it can be seen as a good reaction of what is happening around the world. Among all that happens, the latest trends are most interesting for companies. The latest trends can be analyzed and when indented, reacted to. From a marketing point of view, these latest trends can be used to respond with appropriate activities, like product advertisements. Analyzing tweets can therefore be a goldmine for companies to create an advantage over competitors.

One interesting group is tweets expressing sentiments about products, brands or services. These messages contain an opinion about a specific subject. The sentiment of this opinion can be classified in different categories. An obvious example of three categories is positive, neutral and negative.

The whole process of identifying and extracting subjective information from raw data is known as sentiment analysis. The sentiment analysis researched is closely related to the field of natural language processing. Natural language processing tries to close the gap between human and machine, by extracting useful information from natural language messages.

In data analysis, algorithms have been developed which can be used to analyze data, with the goal to extract useful information. Some widely used classification algorithms from the literature are Naive Bayes and Support Vector Machines. In this project, Naïve Bayes algorithm is used to assign sentiment (positive, neutral or negative) for a tweet.

**1.1 Definitions**

**Sentiment Analysis:**

* According to Wilson, Weber and Homann, "Sentiment analysis is the task of identifying positive and negative opinions, emotions, and evaluations”.
* According to Liu [1] "Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text."

**Sentiment:**

* According to WordNet a sentiment is “A personal belief or judgment that is not founded on proof or certainty”
* According to Wikipedia a sentiment is "An opinion is a subjective statement or thought about an issue or topic, and is the result of emotion or interpretation of facts."

**1.2 Literature survey**

General background of the different related technologies that are involved is discussed in this section, namely sentiment analysis, Twitter and complex event processing.

**1.2.1 Sentiment Analysis**

Sentiment analysis involves several research fields: - natural language processing, computational linguistics and text analysis. It refers to the extraction of subjective information from raw data, often in text form. However, also other media types could contain subjective data, like images, sounds and videos but these types are less studied. In accordance, in all media types different kinds of sentiments exist. The sentiment can refer to opinions or emotions, even though these two types are related there is an evident difference. In sentiment analysis based on opinions, a distinction is made between positive, negative and neutral opinions. The sentiment analysis that is considered in this project is based on opinions and is often referred in literature as opinion mining. Sentiment analysis aims to determine the attitude of the opinion holder with respect to a subject. Other applications try to determine the overall sentiment of a document. Sentiment analysis can be difficult. For example, a text can contain more than one opinion about the same object or about several objects.

Opinion = (oj,fjk,ooijkl,hi,tl)

Where oj is particular object, fjk is feature k of object oj, hi is an opinion holder, tl is the time and ooijkl the actual opinion. Determining the actual polarity of some sentence is the most difficult task of the five properties of quintuple [2]. This sentiment is subjective because different people have different mental scale for what they consider to be a strong or a weak opinion. Therefore it can occur that a sentence is labeled as positive to somebody and neutral as by somebody else.

**1.2.2 Twitter**

Twitter, a micro blogging website which is nowadays familiar to most people, has made a great development in popularity and usage in the last couple of years

The information is contained in the messages (tweets), which have a maximum length of 140 characters. This limited number causes creative people to use acronyms and abbreviations to enlarge the expressibility of their message. Those acronyms lead to a broader dictionary of words, but also make it harder to analyze the tweets, since they create a broader feature space.

Twitter offers a special orthography that includes special features, hash-tags, user mentions and retweets. The "#"-hash-tags are integrated in Twitter, and are used to categorize tweets. In basic usage this categorization is based on subject and topics. But these hash-tags can also be used to add an opposite direction of the tweet, like sarcasm or irony. Such tags can reverse the polarity of the message. The most used hash-tags at the moment are summarized in the trending topics.

With a "@" symbol tweets can be directed to another user. Normally, the tweets are posted in public, or to a restricted group. The prefix "@" with a username directs the message to a specific user. The other user is aware of this directed message, and can respond to it. Thus, conversations can arise by mentioning other user in tweets.

Another Twitter term is the retweet (RT), which is used to show the content of a tweet posted by another user. Users post retweets to note that the original message is interesting enough to send to their followers. An interesting question is whether one should include retweets into your sentiment analysis, since it is actually a repetition of a tweet.

In the line of this research, emoticons are interesting, because they state the mood of a user. This mood is in some cases related and relevant for the sentiment of the message. Smiling and sad emoticons give a good indication of the sentiment; however other emoticons like confused or embarrassed are less informative. Therefore only a part of the emoticons could be useful for sentiment classification.

The most difficult aspect is the overall freedom, because Twitter does not have a protocol about how to use it. This includes spelling mistakes, domain specific content and acronyms. Summarized, the Twitter data lacks a well- defined structure. It is a great challenge to create applications which use Twitter data and accurately determine the sentiment.

**1.2.3 Machine learning:**

Machine Learning techniques can learn normal and anomalous patterns from training data and generate classifiers, which can be used to capture characteristics of interest. In general, the input data to classifiers is an extremely large set of features, but not all of features are relevant to the classes to be classified. Hence, the learner must generalize from the given examples in order to produce a useful output in new cases.

**1.2.3.1 Naïve Bayes**:

Naïve Bayes Classifier is a probabilistic classifier based on applying Bayes’ theorem with strong independence assumption that: the presence of one feature in a class does not depend on the presence or absence of another feature.

The features or also known as attributes are the characterized values to describe an instance. Individual instance is defined by its value on a fixed, predefined set of features or attributes [5]. For example, in the text classification problem, the features can be extracted from words in a document.

The independence assumption does not hold in real texts because of the grammatical relation between words in the sentence.

Naive Bayes [3] is a simple model which works well on text categorization. We use a

multinomial Naive Bayes model. Class c\* is assigned to tweet d, where

|  |  |  |
| --- | --- | --- |
|  | ……………………………... | (1.1) |
|  | ……………………………... | (1.2) |

In the equation (1.2), f represents a feature and ni(d) represents the count of feature fi found in tweet d. There are a total of m features. Parameters P(c) and P (fjc) are obtained through maximum likelihood estimates, and Laplace add-1 smoothing feature is utilized for unseen features [3].

With this algorithm, we consider two representations of a document:

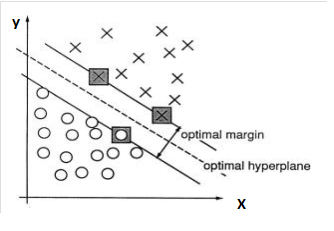
* Naïve Bayes Binary Model (NBB): only presence or absence of words is considered [4].
* Naïve Bayes Multinomial Model (NBM): multiple occurrences of words are considered [4].

For instance, the sentence “my brother is a teacher and my sister is a doctor” is represented as vector of words in two models as below:

* NBB: (my, brother, is, a, teacher, and, sister, doctor)
* NBM: (my, brother, is, a, teacher, and, my, sister, is, a, doctor)

**1.2.3.2 Support Vector Machine**:

Another algorithm for solving the text classification problem is Support Vector Machine (SVM) introduced by [6]. The idea of this algorithm is to consider each document as a point in the document space and to find the appropriate hyper plane to separate the documents into two classes. Figure 1.1 depicts a sample view of the algorithm and documents belong to two classes and the hyper plane which separates them. The x and y are the co-ordinates of two dimensional space [6].



**Figure 1.1 Support Vector Machine illustrations**

However, text classification problem involves with not only two classes but also multiple classes. So the algorithm needs to be extended. There are several works done with the extension of SVM [7]. Two simple approaches are:

* One against all: assume that there are only two classes, one class v/s other classes
* Pair wise classification: one class against one other class and aggregate the results

**1.2.4 Related work**

Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification [8] to learning the polarity of words and phrases [9] [10]. Classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis [11][12]. However, the informal and specialized language used in tweets, as well as the very nature of the micro blogging domain make Twitter sentiment analysis a very different task. It’s an open question of how well the features and techniques used on more well-formed data will transfer to the micro blogging domain. Just in the past year, there have been a number of papers looking at Twitter sentiment and buzz [14]. Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to micro blogging (e.g., emoticons) are also common. Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers depends on emoticons for defining their training data exploit existing Twitter sentiment sites for collecting training data. In [15] also use hash tags for creating training data, but the limitation is their experiments is confined to only sentiment/non-sentiment classification, rather than 3-way polarity classification. Table 1.1 explains the resources and methods used, corpus and results of previous work

**Table 1.1 Related works with result and used methods, resources**

|  |  |  |
| --- | --- | --- |
| **Resources and used methods** | **Corpus** | **Result** |
| AFINN, SentiWordnet, Emoticons, Naive Bayes [16] | 3698 collected tweets after San Bruno event | 0.96 Precision  0.94 Recall |
| Support Vector Machine, Emoticon recognition [17] | Stanford Twitter Sentiment Data | 75.39% accuracy (2 classes)  60.83% accuracy (3 classes) |
| Support Vector Machine, Emoticon recognition [18] | Commercial source, manual elimination | 75.39% accuracy (2 classes)  60.83% accuracy (3 classes) |
| BoosTexter (AdaBoost.MH),n-grams, MPQA(Multi-Perspective Question and Answering) subjectivity lexicon, Internet Lingo Dictionary (emoticons, abbreviations) [19] | HASH(Edinburg)  EMOT  ISIEVE | 75% accuracy (3 classes) |
| Support Vector Machine, Content features, Sentiment Lexicon Features, Target-dependent features, Graph-based optimization [20] | Tweets with specific target {Obama, Google, iPad, Lakers, Lady Gaga} | 85.6% accuracy (2 classes)  68.3% accuracy (3 classes) |

**Table 1.1 Related works with result and used methods, resources**

**1.3 Motivation**

Today micro blogging is become one of the popular communication tools on the Internet. Twitter is an important micro blogging site as millions of people share opinions on different aspects of life every day in the form of tweets. What other people think about a subject has always been a part in decision making; some of the examples are given below

1) “Is it worth to buy this car?”

2) “Which party has a better chance of winning an election?”

3) “Do you think he is a good doctor?”

People tweet on various subjects, for example in the above question on “Is it worth to buy this car?”, numerous tweets can be found which features this car, analyzing for sentiments in these tweets help us determine the popularity of the car and its overall success in the market. Hence sentiment analysis of tweets relating to a particular subject help in determining the popularity of that subject and in turn can be used to assist people in making good decisions.

**1.4 Problem Statement**

To design a dynamic web application that would give the classification of Twitter data as positive, negative and neutral in an easy and efficient way and would provide the best results for the given keyword, the application has to work for the problems listed below-

* The first problem is ungrammatical and unstructured text. Since the message size in Twitter is restricted to 140 characters, Twitter users may have a option to use slangs, abbreviations, or emoticons to shorten the text. It leads to unusual messages.
* The second problem is the fact that tweet messages are not always correct. During fast typing, or using mobile phones as input device, user may have mistyped text and make the analysis step harder.
* The third problem is ambiguity. Due to the less amount of information, it is difficult to identify the objects of interest. For example: “Sony” can either be a laptop brand or a name.
* The fourth problem concerns it is very difficult to find human emotions because they are diverse in nature.

**1.5 Objective**

The main objective of the project is as follows:

* As Twitter data consist of a lot of ambiguities preprocessing of the tweets is done in order to remove retweets, urls, and slang words.
* Feature reduction which increases the efficiency of algorithm.
* Sentiment analysis of data using Naïve Bayes machine learning technique which classifies the data as positive, negative and neutral.
* Evaluation of efficiency of the application for Naïve Bayes algorithm.

**1.6 Scope**

Twitter is a massive medium where people post millions of tweets a day. The English language is the original Twitter language and used most often. So the domain of our project is restricted to English language. Analysis of twitter sentiments help us gain an understanding about what people think about a subject. The sentiment analyzer can be used in the following areas.

* Determining the popularity of product released by a company.
* Predict result of an election.
* Know the popularity of a person.

**1.7 Methodology**

In this project, analysis of data requires the user keyword. Based on the query it retrieves data from Twitter and then it is analyzed by using machine learning technique. Naïve Bayes algorithm is used in this case. The result obtained from the algorithm along with the graphs which help the user understand the sentiment of the data. Hence, the representation of data becomes very important. Having understood this requirement in the early phase of the project, the adopted methodology will accomplish the objectives in a neat and intuitive way. A GUI was developed in the form of Java Server Pages and the back end was coded in Java which helped to exploit the object oriented paradigm in design of algorithms.

**1.8 Organization of report**.

Chapter 2 discusses the software requirements specifications considering product perspective, functional requirements, and software and hardware requirements.

Chapter 3 discusses a high level design of the tool being developed. The data flow diagrams are discussed showing various levels (0, 1, and 2).

Chapter 5 gives the implementation details discussing the programming language selection, coding conventions used. A detailed description of the modules present in this project is presented under this chapter.

Chapter 6 discusses the testing strategies used, testing environment, and various test results are shown.

Chapter 7 explains the results of experiment and performance analysis of various modules of the project.

Chapter 8 summarizes the entire project, stating its limitations and puts forth the possible future enhancements for the project.