**Social Media Sentimental Analysis for Facebook**

Report submitted for the fulfilment of the requirements for the degree of Bachelor of Technology in

**Information Technology**

Under the guidance of

**Mr. Aditya S K**

Submitted by

**Sumukh V Bharadwaj, AMC college of Engineering.**

**Mahesh Kumar S, BNMIT**



**Tequed Labs** - No 10, ANJANEYA NAGAR BSK 3RD STAGE, Bangalore - 560085

**ACKNOLEDGEMENT**

We are greatly indebted to Tequed Labs which has accepted us for the Doctoral Program and provided us with an excellent opportunity to carry-out the present research work.

We are grateful to our guide, mentor, Mr. Aditya S K for having guided us throughout the research span of time and for providing his constructive criticism which made me bring out the best in us. We would also like to thank Sir for being approachable at any point of time without considering his own precious personal time.

We express our sincere gratitude to our family and friends for their whole-hearted support. We would like to convey our thanks to everyone who has been influential and supportive in this research work.

Place: Bengaluru

Date: Sumukh V Bharadwaj

Mahesh Kumar S

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| CHAPTER  NO. | PARTICULARS | PAGE  NO. |
| 1 | Acknowledgement | 2 |
| 2 | Introduction | 4 |
| 3 | Problem statement and its description | 4 |
| 4 | Proposed solution | 5 |
| 5 | Literature survey | 6-19 |
| 6 | Requirement specification |  |
| 7 | Methodology |  |
| 8 | Testing |  |
| 9 | Future Enhancements |  |
| 10 | Reference |  |
| 11 | Appendix |  |

**INTRODUCTION**

**Facebook** (sometimes shortened to **FB**) is a social networking service and website started in February 2004. It was built by Mark Zuckerberg. It is owned by Facebook, Inc. As of September 2012, Facebook has over one billion active users. Users may make a personal profile, add other users as friends, and send messages. Facebook users must register before using the site. The name of the service comes from the name for the book given to students at the start of the school year by some universities in the United States. These books help students get to know each other better. Facebook allows any users who are at least 13 years old to become users of the website.

The growing expansion of contents on the Web, provides a huge collection of textual resources. People share their experiences, opinions or simply talk just about whatever concerns them online. The large amount of available data attracts system developers, studying on automatic mining and analysis. The primary and underlying idea is that the fact of knowing how people feel about certain topic can be considered as classification task. People’s feelings can be positive, negative or neutral. A sentiment is often represented in subtle or complex ways in a text. An online user can use a diverse range of other techniques to express his or her emotions. Apart from this, she/he may mix objective and subjective information about a certain topic. On top of that, data gathered from the World Wide Web often contain a lot of noise. Indeed, the task of automatic sentiment recognition in online text becomes more difficult for all the aforementioned reasons. Hence, we present how sentimental analysis can assist language learning.

**PROBLEM STATEMENT**

Develop a Sentiment Analyzer to analyse user Facebook or WhatsApp messages and to analyse whether the user is in depression or not, if found, info regarding the same should be notified to friends or parents.

**PROPOSED SOLUTION**

We have developed an algorithm using python for sentiment analysis. Python has inbuilt packages for Natural language processing needs and text recognition. Using these packages, the sentiments of any text can be found, in the means of polarity. This is a purely real value and is rational.

Since is Facebook is a much secured space and utilizes end to end encryption, messages can be downloaded in .json format and be subjected to sentiment analysis. Python provides a easy way to handle .json files using a json parser, which is pre-installed. .json formatted files are very easy to work on and data can be extracted in a simpler, hassle-free way.

**LITRATURE SURVEY**

Conducting serious research or making every-day decisions by us often look for other people's opinions. We consult political discussion forums when casting a political vote, read consumer reports when buying appliances, ask friends to recommend a restaurant for the evening. And now the Internet has made it possible to out the opinions of millions of people on everything from latest gadgets to political philosophies. Social media now commands over 22% of the world's total time spent online1 with 65% of adult internet users using some kind of social networking site2. The Internet is increasingly both the forum for discussion and source of information for a growing number of people. As a response to the growing availability of informal, opinionated texts like blog posts and product review websites, Sentiment Analysis has sprung up in the past decade to address the question what do people feel about a certain topic, Bringing together researchers in computer science, computational linguistics, data mining, psychology, and even sociology, sentiment analysis expands the traditional fact-based text analysis to enable opinion-oriented information systems.

• [http://blog.nielsen.com/nielsenwire/global/social-media- accounts-for22-percent-oftime-online/](http://blog.nielsen.com/nielsenwire/global/social-media-%20accounts-for22-percent-oftime-online/)

• <http://pewinternet.org/Reports/2011/Social-Networking-Sites.aspx>

Sentiment analysis (SA) is the field of study that analyses people‘s sentiment and opinions from written (and less often also spoken) language. It can be performed at the document level, the message/sentence level or even the aspect/feature level by them. A popular strategy to deal with the task is to follow a two stage approach. During the first stage, subjectivity detection, a text is classified as subjective if it expresses sentiment, or as objective if it does not. During the second stage, polarity detection, subjective texts are further classified as positive, negative, neutral or sometimes conflict. In some cases the intensity (e.g., strong, mild, weak) of the sentiment is also considered. The classification of texts using this taxonomy has been very popular for the last ten years (Liu, 2012; Pang and Lee, 2005; Tsytsarau and Palpanas, 2012). However, it is not rare for neutral texts to be considered as objective. The marketing research literature purports two basic types of research methods: quantitative and qualitative (Newman, 2011). Quantitative research methods are generally used when researchers are interested in verifying research hypotheses. The research design focuses on collecting data from a large sample of respondents from a defined population, and relies on statistical, mathematical and computational techniques for data analysis (Given, 2008). But, quantitative research is criticized as a rigid approach that ignores inherent subjectivity of human social interactions (Holstein and Gubrium, 1995). On the other hand, qualitative research recognizes multiple realities of human social environment, and is used to discover attitudes, beliefs and emotions on identified phenomenon. But, like quantitative research, qualitative research is also criticized. Opponents describe this approach as a subjective, non-scientific method that lacks structural coherence (Poggenpoel and Myburgh, 2005). Despite the ongoing debate, recent development in research methodologies suggest that the two approaches should be integrated in comprehensive research designs in order to improve research rigor and address several of the epistemological and methodological criticisms (Kelle, 2006; Olsen, 2004). This article contributes to the argument for pluralism in research design by demonstrating how sentiment analysis can be used as a complementary research technique. The paper presents a unique view on the topic of sentiment analysis in social science research by showing how marketers and by extension all stakeholders in the social sciences can benefit from the technique flexibility and scientific rigor. The paper will highlight among other things, the published literature on the topic of sentiment analysis, sentiment analysis methodology, uses of sentiment analysis and the role of sentiment analysis in marketing research. Leaning Approach: With the rise of interconnectivity in our world with the different networks we have and with the amount of information shared, it is becoming highly important for harnessing this information on the web for various reasons. Based on the information collected applications such as market and stock predictions can be put into use. Especially this project focuses on its purpose in industries which are releasing their new products on the market will be eying on how the public responds in order to improve the relation between them and their customers. The data can be analyzed to study the nature of the market which then can be given as a feedback for the desired industry. One merit is that it can be done for any type of product as long as it is on the web.

It can open an area of research in solving the specified problem. There could be different approaches to it. The one and foremost most method which this project also happens to address is putting into use artificial intelligence and machine learning techniques. Sentiment analysis and different clustering and categorizing algorithms such as Bayesian methods and term frequencyinverse document frequency methods are well established methods and widely used. The main purpose of this project is to collect data using an ownbuilt web crawler, which focuses on the compromise between data quantity and quality, along with different APIs and also process the data from

**SENTIMENT ANALYSIS:**

Origins of sentiment analysis are rooted in the disciplines of psychology, sociology and anthropology and flow from the theory of affective stance and appraisal theory which focus on emotions in shaping cognitions. Emotions are feelings generated from both conscious and unconscious processing. An emotional assessment of a situation is a general evaluation of that situation (whether positive or negative) that manifest in mental and bodily responses. The role of emotions in marketing is not new. To the marketer, customer emotions are indirect motivators of purchase behaviour. It shapes brand saliency, influences attitudes, beliefs, opinions and perceptions. Links have already been established between emotions and strong brands (Aaker and Keller, 1990, Morrison and Crane, 2007); emotions and consumption; and emotions on product evaluations (Mano and Oliver, 1993). Sentiment analysis is also not new to market research. Marketers have been analyzing sentiments using old fashion customer comment cards, surveys, interviews and focus groups. Although some of the tools can be adapted to take advantages of the internet interactive environment, their uses 29 are subjected to researcher presence and small sample sizes. Sentiment analysis addresses these problems by systematically collecting and analyzing online sentiments from a very large sample of customers in real time. We conceptualize online sentiments as human convictions or emotions expressed on the internet. It is an attitude towards a situation, event or object, usually expressed through a variety of online media alternatives, with the most popular being social network sites.

Sentiment classification the fourth stage is polarity classification which classifies each subjective sentence in the textual dataset into classification groups. Usually these groups are represented on two extreme points on a continuum (positive, negative; good, bad; like, dislike). However, classification can also involve multiple points similar to the star ratings used by hotels, restaurants and retailers. A wide variety of machine learning techniques are used in binary and polar classification. Machine learning is linked to the field of artificial intelligence and aims at building computational models from past experiences and observation. It fundamentally promotes the use of computer programming to learn and understand fundamentals a particular data set and then use that knowledge acquired to predict or optimize some future criterion. The general objective is to generate a predictive function capable of predicting a target outcome - y (dependent variable) using predefined input criteria or attributes - x (Gama and Carvalho, 2009). When the target is known, this type of learning is called ―supervised learning‖.

Using a supervised leaning approach in sentiment analysis requires training document of textual content or a data corpus, which serves as a preparation document for classification learning. The three basic functions available for classification includes: Naive Bayes (NB), Support Vector Machines (SVM) and MaximumEntropy (ME). A Naive Bayes classifier is a probabilistic classifier based on applying Bayes‘ theorem assuming that features are independent given the class label. This classifier is constructed based on the frequency of occurrence of each feature per class in the training data set. Support vector machines are based on the statistical learning theory (Vapnik, 1995). Binary classifiers show high generalization capability by looking for a 30 hyperplane that maximizes the separation margin between observations from different classes. The use of kernels allows their use for nonlinear problems. Under ME a number of models are constructed where each feature correspond to a constraint on the model. The model with the maximum entropy over all models is selected for classification.

Fairly sophisticated when it comes to understanding the complex meanings beneath the spoken or written word by them. For example, we can tell that a statement like, ―My car had a flat. Brilliant!‖ is sarcastic, not actually brilliant. And with the help of machine learning, computers are beginning to get better at reading between the lines of our tweets, Facebook updates, and email messages, resulting in a new kind of analytics: sentiment analysis. Sentiment analysis, also known as opinion mining, seeks to determine the attitude of an individual or group regarding a particular topic or overall context be it a judgment, evaluation, or emotional reaction from text, video, or audio data. For example, Expedia in Canada used sentiment analysis to determine that the music accompanying one of their commercials was receiving an overwhelmingly negative response online, and they were able to respond to that sentiment appropriately by releasing a new version of the commercial in which the offending violin was abruptly smashed. Say you have a lot of text data from your customers originating from emails, surveys, social media posts, etc. There are several hundred thousand words in the English language. Some are neutral in terms of emotional import, but others have a distinctly positive or negative connotation. This polarity of sentiment can be applied to your customer text to establish what your customers, as a stakeholder group, really think of you. There are number of software tools that can help you to measure text sentiment around your product or service. Facebook post, for example, allows you to separate the positive post about your company, brand, product, or service from the negative and neutral tweets so you can see how well you are doing in the Facebook post. People have long known that surveys and focus groups aren‘t necessarily indicative of broader sentiment. The people who choose to respond to a survey may be the ones who have the most to complain about or the most to praise, but not the middle-of-the-road customers. People brought in for a focus group may 31 alter their opinions based on what they think the company wants to hear. With something like Facebook post analysis, however, you‘re getting the unfiltered opinions of millions of users, not a dozen people sitting in a white room. Sentiment analysis can help you to gauge opinion, which, in turn, can guide strategy and help decision making. In the current business landscape, it‘s increasingly important that we know what our customers, competitors, and employees think about the business, products, and brand. And sentiment analytics can help us do that – often relatively inexpensively. More than Market Research: The technology also is being put to good use outside the marketing and sales arenas. To predict with text-based sentiment analysis which women were at risk of postnatal depression just by analysing their Facebook posts. The research focused on verbal cues that the mother would use weeks before giving birth. Those who struggle with motherhood tended to use words that hinted at an underlying anxiety and unhappiness. There was more negativity in the language used, with an increase in words such as disappointed, miserable, and hate, as well as an increase in the use of ―I‖ – indicating a disconnection from the ―we‖ of impending parenthood.

Qualitative research involves finding out what people think, and how they feel - or at any rate, what they say they think and how they say they feel. This kind of information is subjective. It involves feelings and impressions, rather than numbers. On the other hand, quantitative research focuses on measuring an objective fact. Key to conducting quantitative research is definition of variables of interest and to a large extent a sense of detachment in the data collection by the researcher. Quantitative research analyses data using statistics and relies on large samples to make generalized statements. 15 The relationship between quantitative and qualitative research has never been a smooth and easy one. In fact, there is a heated debate among scientists as to the scientific validity of qualitative research in promoting the advancement of scientific thought. Proponents of qualitative research challenge the credibility of quantitative research claiming that the focus was on merely reinforcing and validating current paradigms rather than discovering new thought. Although the methodological debate continues, a new trend has emerged in research today - the mixed method research design or plural research designs. Plural 32 research design combines both qualitative and quantitative research methods in market studies and is becoming quite a fashion in social science research. It views qualitative/quantitative techniques are merely tools used in understanding the world we live in. Both tools are united by a shared commitment in knowledge creation, knowledge dissemination and to a rigorous, conscientious research process. Researchers are encouraged to implement plurality through triangulation which involves both qualitative and quantitative approach to data collection and design. However, a recent survey of research papers discovered several methodological deficiencies with triangulated designs given the absence of systematic and scientific guidelines (Kelly, 2006). Triangulation also can be impractical to some research situations given the high research cost of multiple data collection and the time delays in data collection and data analysis. Sentiment analysis is a useful tool to address triangulation challenges in an online environment. When integrated with qualitative research, sentiment analysis can be used as a tool that promotes rigor and structure to an otherwise flexible and subjective data collection and data analysis process. Alternatively, if integrated with quantitative research sentiment analysis facilitates a deep rich insight into unsolicited opinions and emotions, thus facilitating a more meaningful understanding of any phenomenon. By employing machine learning techniques, sentiment analysis presents an opportunity to lend a systematic approach to mixed method design. We argue that through sentiment analysis, a marketer is presented with a rich option to procure meaningful and insightful feedback into customer feelings, thoughts, opinions and sentiments in real time. Sentiment analysis provides a faster, simpler and less expensive alternative to traditional qualitative market research techniques like observations, interviews and even ethnography as well as provides information in real time. At the same time, it offers the advantages of traditional quantitative methods including measurability and objectivity. Data is also collected in a manner that is entirely unobtrusive as compared to methods used in both qualitative and quantitative. Sentiment analysis provides an opportunity for marketers to collect data on customers in their natural cyber environment, without the presence of the researcher being felt. 33 Therefore this method eliminates the problem of people reacting differently when they know their responses are being collected.

Why is Sentiment Analysis important when Analysing Social Media: Social media sentiment analysis can be an excellent source of information and can provide insights that can:

• Determine marketing strategy

• Improve campaign success

• Improve product messaging

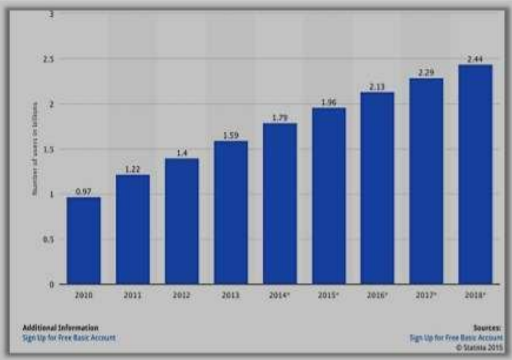
• Improve customer service

• Generate leads

In a nutshell, if done properly, social media sentiment analysis can improve your bottom line. However, if you are making decisions using incorrect sentiment analysis data, the results can be catastrophic. Most social media analysis vendors will admit (if you push them hard enough) that their sentiment analysis algorithm will be, at best, 50-60% accurate. The study of sentiment analysis, if done properly, is exceptionally complex and is actually a field of study, not just a feature in a social media tool. I should probably be clear at this point, that the objective of this blog is not to discuss the nuances and detail of sentiment analysis. In fact, quite the opposite, To simplify this very complex topic so that you can use the information when deciding on the social media tool or services that you need by us. You probably recognise the complexity when you realise the number of times people misinterpret conversations or miscomprehend the written word. Also, elements such as sarcasm and jargon make it even more difficult to determine meaning from words. So how do we simplify this topic so that we can evaluate vendor tools There are a number of factors to take into consideration.

**Types of Sentiment Analysis**

Need to understand the methods that social media vendors use to determine sentiment by us. There are many types of sentiment analysis. However, for the purposes of this article, concentrate on three by us. Manual processing: Human interpretation of sentiment is definitely the most mature and accurate judge of sentiment. However, it still isn't 100% accurate. Very few vendors still use this process without the additional use of a tool. This is due to the prolific growth of social media. According to Seth Grimes, social is the fastest growing source of enterprise analytical data



Number of social network users worldwide from 2010 to 2018 (in billions): Therefore, if you are going to use social media to determine sentiment, it is becoming less practical to use human processing and more likely you will need to automate the process.

**Keyword processing:**

Keyword processing algorithms assign a degree of positivity or negativity to an individual word, then it gives and overall percentage score to the post. For example, positive words, great, like, love or negative words: terrible, dislike

The advantages of this method are that it is very fast, predictable and cheap to implement and run. However, there are numerous disadvantages including dealing with double negatives or positives or different meanings of words, for example, the use of a word such as 'sick' (to mean either "ill" or to mean "awesome"). Not to mention, different researchers may assign difference percentages of positive or negative to a word. More often the issue is that it does not deal with multiple word/context issues or non-adjective words.

Natural Language Processing (NLP also called: text analytics, data mining, computational linguistics): NLP refers to computer systems that process human language in terms of its meaning. NLP understands that several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas. NLP works by analysing language for it's meaning, NLP systems are used for in a number of areas such as converting speech to text, language translation, and grammar checks. It can be likened to programming an algorithm to interpret the English language (or any language for the matter) with the rules that you were taught in English class. Although NLP may seem to be far superior to keyword processing, it still has its limitations. Sarcasm a well known Australian trait, is very difficult to detect using NLP as is hyperbole and social media acronyms or social jargon such as:

• Youturn: To follow another person on social media with the intention of following

• Wallflower: A person who regularly consumes the social media of others but never posts

• Face Crawling: Begging for Facebook likes, online or offline

• Hash-Browning: The excessive use of hashtags within a single post People express opinions in complex ways for example, Sometimes even human interpretation can be hard to determine.

**TOOLS AND WORKS IN SENTIMENT ANALYSIS**

According to Pang and Lee (2008), researchers have found ways to avoid the use of manual annotation by utilizing existing online textual content generated from sites such as Epinion, Amazon, Rotten Tomatoes, Twitter, Facebook. Several sentiment search engines exist where users run typical queries on any topic of interest, and generate text results. Usually the results are coded and categorized into two or three polar categories. Some examples currently available are: Twitrratr: www.twitrratr.com Sentiment 140 - http://www.sentiment140.com Tweet feel www.tweetfeel.com and Opinmind www.opinmind.com. Social Mention www.socialmention.com Sentiment search engines make sentiment analysis quite easy. But, the online reviews on sites like Amazon and Opinion have been found to be skewed towards the positive which raises questions on validity and reliability of sentiment classification. However, Pang and Lee (2008) admit that although the content might be skewed, the validity of the process is acclaimed. Another tool in sentiment analysis is word lists or annotated databases which categorize words based on their emotions for example -attractive (positive valance) or aversive (negative valance). Some examples include: ANEW, General Inquirer and LIWC. Other tools include sentiment analysis programs that are specifically designed to categorize short textural documents. One example is sent strength. Machine Learning Algorithms: Machine learning is a scientific discipline that deals with the construction and study of algorithms that can learn from data (Kovahi and Provost, 1998). Such algorithms operate by building a model based on inputs (Bishop, 2006) and using that to make predictions or decisions, rather than following only explicitly programmed instructions.

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are:

**Supervised learning –**

The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

**Unsupervised learning –**

No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.

**Reinforcement learning –**

A computer program interacts with a dynamic environment in which it must perform a certain goal, without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent. Among the most common Machine Learning approaches are below:

a) Decision Tree Learning

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value.

b) Artificial Neural Networks

An artificial neural network (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is inspired by the structure and functional aspects of Biological neural networks. Computations are structured in terms of an interconnected group of artificial neurons, processing information using a connectionist approach to computation. Modern neural networks are non-linear statistical data modelling tools. They are usually used to model complex relationships between inputs and outputs, to find patterns in data, or to capture the statistical structure in an unknown joint probability distribution between observed variables.

c) Bayesian Networks

A Bayesian network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Efficient algorithms exist that perform inference and learning.

d) Clustering

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to some pre-designated criterion or criteria, while observations drawn from different clusters are dissimilar. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis.

e) Classification

Classification is the task of assigning a label to an input. The general class of classification problems includes many kinds of input besides text. To cite a few well-known examples, there are market basket systems, anomaly detection systems, and vision systems. A market basket system tries to detect buying patterns from a buyer‘s purchasing records and other information about them and their friends; an anomaly detection system tries to detect deviation from normal event patterns that signal trouble, for example in a computer security setting, or in a credit card fraud detection setting. A vision system is a system that recognizes objects or events.

f) Genetic Algorithms

A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection, and uses methods such as mutation and crossover to generate new genotype in the hope of finding good solutions to a given problem. In machine learning, genetic algorithms found some uses in the 1980s and 1990s

**Classification**

As described in (e) above, classification is the task of assigning a label to an input. There are two basic steps to using a classifier: training and classification. Training is the iterative process of taking content that is known to belong to specified classes and creating a classifier on the basis of that known content. Classification is a onetime process of taking a classifier built with such a training content set and running it on unknown content to determine class membership for the unknown content. There are two main approaches for classification: supervised and unsupervised classification. In supervised classification, the classifier is trained on labelled examples that are similar to the test examples, whereas unsupervised learning techniques assign labels based only on internal differences (distances) between the data points. In this classificationapproach each sentence is considered independent from other sentences (Yessenov and Misailovic, 2009). Text classification is a special kind of classification problem. There are many practical applications of text classification.

**Respect to some topic or the overall contextual polarity of a document.**

Categorizing news feed topics: – classifying text according to topics.First there is a set of labelled training documents, which means we have, not just the documents, but some indication of what class they belong to from a small set of classes. From each document a vector of features is extracted. The features are the representations of the documents which the learner uses to try to draw generalizations about how to predict classes. From the feature representations of the training documents and their labels, the learner produces a classifier. This phase is called the training phase. The classifier produced in the training phase can be used to classify new, unseen documents. To do this, features are extracted from the new document; the features are passed to classifier, and a classification decision (expected label) is produced.

**Text Classification Techniques**

The most common Machine Learning algorithms for sentiment classification are Naïve Bayes, Maximum Entropy and Support Vector Machine. These are described below.

**Naïve Bayes Classifier**

The Naive Bayes classifier is an extremely simple classifier that relies on Bayesian Probability and the assumption that feature probabilities are independent of one another (Vachaspati, P and Wu, C., 2012). In simple terms, a naive bays classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. This assumption is called class conditional independence. For example, a fruit may be considered to be an apple if it is red, round, and about 3" in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features.

**The Maximum Entropy Classifier:**

The Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. Unlike the Naive Bayes classifier discussed in the previous section, the Max Entropy does not assume that the features are conditionally independent of each other. The MaxEnt is based on the Principle of Maximum Entropy and from all the models that fit the training data, selects the one which has the largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.Due to the minimum assumptions that the Maximum Entropy classifier makes, it is regularly used when nothing is known about the prior distributions and when it is unsafe to make any such assumptions. Moreover Maximum Entropy classifier is used when the conditional independence of the features cannot be assumed. This is particularly true in Text Classification problems where the 41 features are usually words which obviously are not independent. The Max Entropy requires more time to train compared to Naive Bayes, primarily due to the optimization problem that needs to be solved in order to estimate the parameters of the model. Nevertheless, after computing these parameters, the method provides robust results and it is competitive in terms of CPU and memory consumption. Sentiment analysis/classification (or opinion mining) is defined as the task of finding the opinions of authors about specific entities. The decision-making process of people is affected by the opinions formed by thought leaders and ordinary people. When a person wants to buy a product online he or she will typically start by searching for reviews and opinions written by other people on the various offerings. Sentiment analysis is one of the hottest research areas in computer science. Over 7,000 articles have been written on the topic. Hundreds of start-ups are developing sentiment analysis solutions and major statistical packages such as SAS and SPSS include dedicated sentiment analysis modules. There is a huge explosion today of ―sentiments‟ available from social media including Twitter, Facebook, message boards, blogs and user forums. These snippets of text are a gold mine for companies and individuals that want to monitor their reputation and get timely feedback about their products and actions. Sentiment analysis offers these organizations the ability to monitor the different social media sites in real time and act accordingly. Marketing managers, PR firms, campaign managers, politicians and even equity investors and online shoppers are the direct beneficiaries of sentiment analysis technology (Feldman, 2013).

**Gaps Identified**

• From the review, it was found that there is no study conducted on social media sentimental analysis in India.

• Found that there was no study conducted on social media sentimental analysis using platform of Facebook.

Research Issues on Social Network Analysis: A number of research issues and challenges facing the realisation of utilising data mining techniques in social network analysis could be identified as follows: Linkage-based and Structural Analysis – This is an analysis of the linkage behaviour of the social network so as to ascertain relevant nodes, links, communities and imminent areas of the network Aggarwal, 2011. Dynamic Analysis and Static Analysis Static analysis such as in bibliographic networks is presumed to be easier to carry out than those in streaming networks. In static analysis, it is presumed that social network changes gradually over time and analysis on the entire network can be done in batch mode. Conversely, dynamic analysis of streaming networks like Facebook and YouTube are very difficult to carry out. Data on these networks are generated at high speed and capacity. Dynamic analysis of these networks are often in the area of interactions between having presented some of the research issues and challenges in social network analysis, the following sections and sub-sections present the overview of different data mining approaches used in analysing social network data. Graph Theoretic: Graph theory is probably the main method in social network analysis in the early history of the social network concept. The approach is applied to social network analysis in order to determine important features of the network such as the nodes and links (for example influencers and the followers). Influencers on social network have been identified as users that have impact on the activities or opinion of other users by way of followership or influence on decision made by other users on the network. Graph theory has proved to be very effective on large-scale datasets (such as social network data). This is because it is capable of bye passing the building of an actual visual representation of the data to run directly on data matrices employed parameterized centrality metric approach to study the network structure and to rank nodes connectivity. Work formed an extension of acentrality approach which measures the number of alleviated paths that exist among nodes by them. A community is a smaller compressed group within a larger network. Community formation is known to be one of the important characteristics of social network sites. Users with similar interest form communities on social network thereby displaying strong sectional structure.

**METHODOLOGY**

This is a very simplified code for sentiment analysis. The program is a single function based code which does both parsing and sentiment analysis.

**The JSON Parser**

JSON parser is a simple framework for manipulating JSON files. It converts any String read to JSON dictionaries where data can be accessed easily. The JSON Parser is used, as the messages are in dictionary form. The Facebook chat will be in the following format:

{

“participants”:[

{

“name” : ”user1”

},

{

“name” : ”user2”

}

]

“messages”:[

{

“sender\_name” : ”username”,

“timestamp\_ms” : 12345676789,

“content” : “Message”, #important

“type” : “Gereric”

},

.

.

.

.

}

The JSON file contains the complete information you need to know about the chat. But the important parts for the job is the “content” key, the participants, the “sender\_name” key. The program uses the above data for analysing the relationship sentiment of the two conversing users.

Analysis is done only for text and sometimes when the sender uses sticker, the content field is absent. So a simple if statement is used to tackle and use only data having content keys.

def cleaner(data):

if 'content' in data.keys(): return True

.

.

if cleaner(i)==True:

msg=i['content']

.

.

**The TextBlob framework**

TextBlob is a framework for natural language text parsing and considering portions of text for analysing. TextBlob is a purely a dealer of String datatypes so it is necessary to ensure the function parameters are Strings.

Like any other parser, TextBlob should store an operateable copy into another variable and use the former to get the sentiment polarity.

tb\_msg=TextBlob(msg)

As the program further proceeds, in order to find the percentage of the messages, the polarity is cumulated and the message count is taken for each participant, by checking the names.

if(i['sender\_name']==person1):

score1=+tb\_msg.sentiment.polarity

count1=+1

else:

score2=+tb\_msg.sentiment.polarity

count2=+1

Finally, when it comes to the analysis and result generation, sentiments are classified as either a good relationship or a bad one, by checking the total score and conditional decisions.

if score2<0: print('Verdict: You don\'t have a good relationship with this person.')

elif score2>0: print('Verdict: You have a good relationship with this person.')

else: pass

**THE TESTING PHASE**

This is a program which deals in non numeric data, as sentiments cannot be numerically evaluated. Keeping this in mind, testing was carried out with the best cases of tests accurate to the current time frame.

**Boundary Value Analysis:**

Boundary value analysis or BVA tests the range of values compatible for the program and also analyses the values returned by the program.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test case No | Person 1 | Person 2 | Score2 | Verdict | Actual Status |
| 1 | FBUser\_1 | Sumukh | 13.5% | You Have a Good relationship | Good relationship |
| 2 | FBUser\_2 | Sumukh | 0% | - | Neutral |
| 3 | FBUser\_3 | Sumukh | -10% | You don’t have a good relationship | Bad Relationship |
| 4 | FBUser\_1 | Mahesh | 0% | - | Neutral |
| 5 | FBUser\_2 | Mahesh | 5% | You Have a Good relationship | Good relationship |
| 6 | FBUser\_3 | Mahesh | -6% | You don’t have a good relationship | Bad Relationship |

**BVA Graph:**

This is the general trend in the outputs of the above test

**Path testing**

This test jots down a path which will indicate the program flow when it is run

Note: Refer the Appendix for the line numbers of the program

|  |  |
| --- | --- |
| Line Number(s) | Node |
| 1-5 | A |
| 6-7 | B |
| 8-15 | C |
| 16 | D |
| 17 | E |
| 18-19 | F |
| 20-22 | G |
| 23-25 | H |
| 27-29 | I |
| 30 | J |
| 31 | K |
| 32 | L |

**Graph:**

**KEY FEATURES**

1. Simple and effective
2. Ease of usability
3. Can be used for any chat backup in JSON format
4. Portable
5. Efficient
6. Consumes less memory

**FUTURE ENHANCEMENTS**

This algorithm is useful for various behavioural analysis and many other applications. Simple as it is, this algorithm can be used in many other applications, like forensic deductions of suspects and the victim, along with behavioural analysis for detective purposes. Another application which is much needed in the present times, that is suicide prevention. Relationships can analysed for predicting depressed moods, which can help in giving precautions, prior warnings, counselling processes etc.

**APPENDIX**

**Code Section:**

1: from textblob import TextBlob

2: import json

3: import sys

4: non\_bmp\_map = dict.fromkeys(range(0x10000, sys.maxunicode + 1), 0xfffd)

5:

6: def cleaner(data):

7: if 'content' in data.keys(): return True

8: str1=open('facebook-sumukhbharadwaj96/messages/inbox/athiramnair\_xujrplfhfw/message\_1.json').read()

9: data=json.loads(str1)

10: persons=data['participants']

11: person1=persons[0]['name']

12: person2=persons[1]['name']

13: count1=0

14: count2=0

15: msgData=data['messages']

16: for i in msgData:

17: if cleaner(i)==True:

18: msg=i['content']

19: tb\_msg=TextBlob(msg)

20: if(i['sender\_name']==person1):

21: score1=+tb\_msg.sentiment.polarity

22: count1=+1

23: else:

24: score2=+tb\_msg.sentiment.polarity

25: count2=+1

27: print('Sentiment Analysis')

28: print('Polarity of '+person1+': '+str((score1/count1)\*100))

29: print('Polarity of '+person2+': '+str((score2/count2)\*100))

30: if score2<0: print('Verdict: You don\'t have a good relationship with this person.')

31: elif score2>0: print('Verdict: You have a good relationship with this person.')

32: else: pass