GROUPING OF SENTIMENTS OF SOCIAL NETWORK

A thesis submitted in partial fulfillment of the requirement for the

Degree of Master of Computer Application

of

Jadavpur University

By

DEBARATI BERA

Registration Number: 137330 of 2016-2017

Examination Roll Number: MCA196019

Under the Guidance of

Dr. Diganta Saha

Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

May, 2019

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled "GROUPING OF SENTIMENTS OF SOCIAL NETWORK" has been satisfactorily completed by Debarati Bera (University Registration No.: 137330 of 2016-17, Examination Roll No.:MCA196019). It is a bonafide piece of work carried out under my guidance and supervision and be accepted in partial fulfillment of the requirement for the Degree of Master of Computer Application , Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, Kolkata.

Dr. Diganta Saha (Thesis Supervisor)
Professor
Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Countersigned

Prof. Mahan Tapas Kundu
Head, Department of Computer Science and Engineering,
Jadavpur University, Kolkata-700032.

Prof. Chiranjib Bhattacharjee
Dean, Faculty of Engineering and Technology,
Jadavpur University, Kolkata-700032.

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled “GROUPING OF SENTIMENTS OF SOCIAL NETWORK” is a bonafide record of work carried out by Debarati Bera in partial fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University during the period of February 2019 to May 2019. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn there in but approve the thesis only for the purpose for which it has been submitted.

____

Signature of Examiner
Date:

____

Signature of Supervisor
Date:

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis entitled "GROUPING OF SENTIMENTS OF SOCIAL NETWORK" contains literature survey and original research work by the undersigned candidate, as part of his Degree of Master of Computer Application.
All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.
I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: DEBARATI BERA
University Registration No. : 137330 of 2016-17
Examination Roll No. : MCA196019

Thesis Title: GROUPING OF SENTIMENTS OF SOCIAL NETWORK

____
Signature
Date:

ACKNOWLEDGEMENT

First and foremost, I would like to start by thanking God Almighty for showering me with the strength, knowledge and potential to embark on this wonderful journey and to persevere and complete the embodied research work satisfactorily.

I am pleased to express my deepest gratitude to my thesis guide, **Dr. Diganta Saha**, Department of Computer Science and Engineering, Jadavpur University, Kolkata for his invaluable guidance, constant encouragement and inspiration during the period of my dissertation.

I am highly indebted to **Jadavpur University** for providing me the opportunity and the required infrastructure to carry on my thesis.

I am thankful to all the teaching and non-teaching staff whose helping hands have smoothed my journey through the period of my research.

Last but not the least; I would like to thank my family members, classmates, seniors and friends for giving me constant encouragement and mental support throughout my work.

Debarati Bera
University Registration No. : 137330 of 2016-19
Examination Roll No. : MCA196019
Master of Computer Application
Department of Computer Science and Engineering
Jadavpur University

Abstract

The aim of this project targets to make a sentiment analysis of tweets and predict the positivity or negativity of the sence of the user. The Sentiment Analysis techniques used here are unsupervised methods. This is used to classify various tweets. Our hypothesis is that we can obtain high accuracy on classifying sentiment in Twitter messages using unsupervised techniques. Sentiment analysis of the tweets determine the polarity and inclination of vast population towards specific topic, item or entity.The sentiment analysis of the emojis allows us to draw several interesting conclusions. It turns out that most of the emojis are positive, especially the most popular ones.The sentiment distribution of the tweets with and without emojis is significantly different. We will propose a system that will analyze tweets about three categories positive, negative and neutral. For this we use unsupervised learning and clustering method. This paper also discuses about problems in sentiment analysis, proposed system, existing system and work flow of proposed system. The main contribution of this project is figure out the subjectivity, polarity and sentiment in the text. Our hypothesis presents an unsupervised learning algorithm for clustering tweets i.e. K-means clustering.

Key words: sentiment, cluster, polarity.

Index

Contents **Page No.**

Chapter-1

Chapter-2

Chapter-3

Chapter-4

Chapter-5

CHAPTER 1:

1.1 Introduction

1.1.1 Introduction to sentiment analysis

The emergence of social media has given web users a venue for expressing and sharing their thoughts and opinion on different topics and events. Twitter, with nearly 600 million users and over 250 million messages per day, has quickly become a gold mine for organizations to monitor their reputation and brand by extracting and analyzing the sentiments of the tweets posted by public about them, their markets, and competitors . Sentiment analysis (also known as opinion mining), which commonly refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials , has become a valuable area of research and has attracted many researchers from both academia and industry. In the literature, sentiment analysis is treated as a machine learning process that aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. Generally, three main phases of data pre-processing, vector space modelling (VSM) and sentiment analysing (learning) are involved.

A basic task 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, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".

1.1.2 Importance of sentiment analysis

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
- Test business KPIs (Key Performance Indicator)
- Generate leads
- Develop product quality
- Crisis management

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.

1.1.3 Sentiment analysis Terminology

In this segment we need the various terms used in the Sentiment Analysis.

Fact: A fact is that which has truly happened or which is really the case.

Opinion: An opinion is a view or judgment formed about something (like Product or movie) not necessarily based on fact or knowledge.

Subjective Sentence: A sentence or a text is a subjective or opinionated if it actually indicates ones feelings or emotions.

Objective Sentence: An objective sentence indicates some facts and known Information about the world. For example: universal truths.

Review: A review is texts that contain a particular combination of words that has opinions of customer a particular item or opinions of viewers for a movie. A review may be subjective or objective or even both.

Item: An individual article or unit, especially one that is part of a list, collection or set.

Known Aspects: Known aspects are default aspects provided by the certain website for which users separately give ratings.

Sentiment: Sentiment is a polarity term that implies to the direction in which a behavior or opinion is expressed. For example, excellent is a sentiment for the attribute camera in the sentence "the camera of the iphone is excellent".

Opinion Polarity: Opinion Polarity or Subjectivity Orientation denotes the polarity expressed by the user or customer or viewer in terms of numerical values. Rating: Most of the people use star ratings for expressing polarity, represented by stars in the range from 5 to 1 which is called ratings.

Polarity: Polarity is a three way orientation scale. In this, a sentiment can be either negative or positive or neutral.

1.1.4 Introduction to clustering

Cluster analysis is the unsupervised process of grouping data instances into relatively similar objects, without prior understanding of the groups structure or class labels . Cluster analysis is one of the traditional topics in the data mining field. It is the first step in the direction of exciting knowledge discovery. Clustering is the procedure of grouping data objects into a set of disjoint classes, called clusters. Now objects within a class have high resemblance to each other in the meantime objects in separate classes are more unlike.

Clustering is a fundamental problem that has neumorous applications in many disciplines. Clustering techniques are used to discover natural groups of data sets and to identify abstract structures that might reside there, without having any background knowledge of the characteristics of the data. They have been used in a variety of areas, including bioinformatics; compuetr vision; data mining; gene expression analysis; image segmentation; information retrieval; machine learning; text mining; signal compression; etc. . Grouping or clustering is a building block in a wide range of application.

The classical k-means clustering algorithm is introduced first with the squared Euclidian distance. In k-Means algorithm, only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroid to members of the data set, choosing medians, choosing an initial center less randomly.The algorithm prefers clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters . K-means has a number of interesting theoretical properties. On the one hand, it partitions the data space into a structure known as a Voronoi diagram. On the other hand, it is conceptually close to nearest neighbor classification, and as such is popular in machine learning. Third, it can be seen as a variation of model based classification .

1.2 Conclusion

Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. The importance of sentiment analysis on social network is increasing day by day. So, it is a good field to work with clustering method.

CHAPTER 2:

2.1 Literature Reviews

2.1.1 Introduction

A literature review discusses published information in a particular subject area, and sometimes information in a particular subject area within a certain time period. It can be an analysis of literature or published sources, on a particular topic.

A literature review is a simple summary of the sources, but it usually has an organizational pattern and combines both summary and synthesis. A summary is a recap of the important information of the source, but a synthesis is a re-organization, or a reshuffling, of that information. It might give a new interpretation of old material or combine new with old interpretations. Or it might trace the intellectual progression of the field, including major debates. And depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant.

2.1.2 Survey of Related Work

Existing works on web text mining and clustering are mainly focused on the different levels like: Web text clustering, Data text mining, Web page information extraction etc. In recent years a lot of work has been done in the field of “Sentiment Analysis on Twitter “by number of researchers. In its early stage it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative only. They proposed either machine learning approach or lexicon-based approach or they may even include combination of both to achieve good accuracy.

Some of literature reviews are:

As research in Indonesia, Bojar who conducted research about the resources of the lexicon for Indonesian sentiment also did the negation handling. By adapting the technique from Das and Chen. handled the negation of sentiment caused by a negation word. Bojar uses negation words such as ‘tidak’, ‘tak’, ‘tanpa’, ‘belum’, and ‘kurang’. The words that occur between the negation words and the first punctuation after the negation word are tagged with ‘NOT_’. Example, there is a sentence: ‘kameranya kurang bagus gambarnya’ became ‘kameranya kurang NOT_bagus NOT_gambarnya’.

Turney et al (2002) used bag-of-words method for sentiment analysis in which the relationships between words was not at all considered and a document is represented as just a collection of words. To determine the sentiment for the whole document, sentiments of every word was determined and those values are united with some aggregation functions.

Preceded by Pang Lee et. al , they classify documents not by topics but by sentiments, e.g. determining whether the review is positive or negative. For negation handling, if a word x follows the negation word then a new feature ‘NOT_x’ created tag every word from x until first punctuation mark. But this method cannot model the scope of negation, because it is heuristically tagging all word until it finds the mark, without concerning with negation words or not. Addition in preprocessing task, mostly the punctuation marks is removed; this is for simplification in preprocessing stage.

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task (Turney, 2002; Pang and Lee, 2004), it has been handled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009).

Parikh and Movassate(2009) implemented two models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model.

Murali Krishna and Durga Bhavani (2010) proposed the use of a renowned method, called Apriori algorithm, for mining the frequent item sets and devised an efficient approach for text clustering based on the frequent item sets. Maheshwari and Agrawal (2010) proposed centroid-based text clustering for preprocessed data, which is a supervised approach to classify a text into a set of predefined classes with relatively low computation and better accuracy📬

Web data clustering researchers Bouras and Tsogkas (2010) proposed an enhanced model based on the standard k-means algorithm using the external information extracted from WordNet hypernyms in a twofold manner: enriching the “bag of words” used prior to the clustering process and assisting the label generation procedure following it.

Pak and Paroubek(2010) proposed a model to classify the tweets as objective, positive and negative. They created a twitter corpus by collecting tweets using Twitter API and automatically annotating those tweets using emoticons. Using that corpus, theydeveloped a sentiment classifier based on the multinomial Naive Bayes method that uses features like N-gram and POS-tags. The training set they used was less efficient since it contains only tweets having emoticons.

Davidov et al.,(2010) proposed a approach to utilize Twitter user-defined hastags in tweets as a classification of sentiment type using punctuation, single words, n-grams and patterns as different feature types, which are then combined into a single feature vector for sentiment classification. They made use of K-Nearest Neighbor strategy to assign sentiment labels by constructing a feature vector for each example in the training and test set.

Qiujun (2010) proposed a new approach to news content extraction using similarity measure based on edit distance to separate the news content from noisy information. Jaiswal (2007) performed the comparison of different clustering methods like K-Means, Vector Space Model (VSM), Latent Semantic Indexing (LSI), and Fuzzy C-Means (FCM) and selected FCM for web clustering.

Bifet and Frank(2010) used Twitter streaming data provided by Firehouse API,which gave all messages from every user which are publicly available in real-time. They experimented multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They arrived at a conclusion that SGD-based model, when used with an appropriate learning rate was the better than the rest used.

Xia et al.(2011) used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In thier work, they used two types of feature sets (Part-of-speech information and Word-relations) and three base classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines). They applied ensemble approaches like fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy.

In 2012, Balakrishnan Gokulakrishn et. al. proposed an approach where a plugged stream of tweets from the Twitter micro blogging webpage are preprocessed and grouped in light of their emotional content as positive, negative and irrelevant; and investigates the execution of different ordering calculations in light of their precision and recall in such cases.

Po-Wei Liang et.al.(2014) used Twitter API to collect twitter data. Their training data falls in three different categories (camera, movie, mobile). The data is labeled as positive, negative and non-opinions. Tweets containing opinions were filtered. Unigram Naive Bayes model was implemented and the Naive Bayes simplifying independence assumption was employed. They also eliminated useless features by using the Mutual Information and Chi square feature extraction method. Finally, the orientation of an tweet is predicted. i.e. positive or negative.

In 2014, Aizhan Bizhanovaet. al. proposed a model for naturally characterizing the opinion of Twitter messages toward item/mark, utilizing emoticons and by enhancing preprocessing steps keeping in mind the end goal to accomplish high exactness.

In 2014,Calvinet. al. proposed a model where sentiment extremity of Twitter surveys are measured utilizing Naïve Bayes classifier strategy. The model demonstrates a promising come about on characterizing the ubiquity in light of consume satisfaction and along these lines characterizing the best supplier to be utilized.

In 2016, Sandip D Mali et. al. proposed another framework called SentiView which a vocabulary based approach for sentiment investigation. They have gotten high accuracy because of preprocessing and expulsion of non-opinion tweets from data.

Pablo et. al. presented variations of Naive Bayes classifiers for detecting polarity of English tweets. Two different variants of Naive Bayes classifiers were built namely Baseline (trained to classify tweets as positive, negative and neutral), and Binary (makes use of a polarity lexicon and classifies as positive and negative. Neutral tweets neglected). The features considered by classifiers were Lemmas (nouns, verbs, adjectives and adverbs), Polarity Lexicons, and Multiword from different sources and Valence Shifters.

In 2016, Sanjana Woonnaet. al. proposed a framework that examinations tweets into three classifications which are positive, negative and neutral utilizing supervised learning approach After the execution, the outcomes demonstrated which viewpoints individuals like or aversion and how feelings on motion pictures changes over a timeframe.

In 2017,Kai Yang et. Al proposed a highly effective hybrid model combining different single models to overcome their weaknesses. They build the sentimental dictionary from exterior data. As single model have many limitations and weakness. That why they build a hybrid model by combing many single approaches to overcome those limitations of single model. The experimental results show that our hybrid model shows very great performance. In hybrid model 2 approaches that are SVM and GDBT (Gradient boosting decision tree) are combined together that are based on stacking approach.

For the negation in sentiment, there are some of the researchers that focus on the impact of the negation in sentiment sentences. A survey conducted by Wieghan et. al , they survey for negation role in sentiment analysis. They state that effective negation model for sentiment analysis usually requires the knowledge of polar expression. Jia et. al. studied the impact of each occurrence of term in a sentence on its polarity and introduced the concept of scope of the term t.

2.2 Conclusion

The work on sentiment analysis has been started earlier. The analysis on sentiment of text helps to recognise the positivity or negativity of the text. It introduced a wide area in analysis of social media text and analysing the sentiment of user and also helps in reviews.

CHAPTER 3:

3.1 Sentiment Analysis Techniques

3.1.1 Academic Prospective of Sentimental Analysis

Sentiment Analysis (SA) or Opinion Mining (OM) – is a discipline that has seen a lot of activity since about 2000.The verge expansion and associability of the social media's and proliferation of social media and its tools (e.g. Twitter, Facebook, LinkedIn, etc.), that has made the accessibility to information about how people feel about things more readily available to the masses. The main fields of research in sentimental analysis are Subjectivity Detection, Sentiment Prediction , Aspect Based Sentiment Summarization, Textual Summarization, Constrictive Viewpoint Summarization , Opinion based-entity ranking , Review Detection, Product Feature Extraction ,Opinion Retrieval .

An important part of our information-gathering behavior has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites, new opportunities and challenges arise as people can, and do, actively use information technologies to seek out and understand the opinions of others.Opinion Mining and Sentiment Analysis covers techniques and approaches that promise to directly enable opinion-oriented information-seeking systems. The focus is on methods that seek to address the new challenges raised by sentiment-aware applications, as compared to those that are already present in more traditional fact-based analysis.

- Subjectivity Detection is about determining if a piece of text actually contains opinions or not (i.e. subjective expression or objective?).

- Sentiment Prediction is specifically about predicting the polarity as it is positive or negative or neutral at the literally level.

- Aspect Based Sentiment Summarization is to provide a sentiment summary of service or product at the feature or aspect level (i.e. start rating or service score).

- Textual Summarization gathers a few informative sentences or phases that summaries the review of the product.

- Constrictive Viewpoint Summarization is to highlight contradiction in opinions.

- Opinion based-entity ranking is task of ranking entities based on opinions.

- Review Detection deals with identify the real comments and about to identifying the fake opinion from reviews.

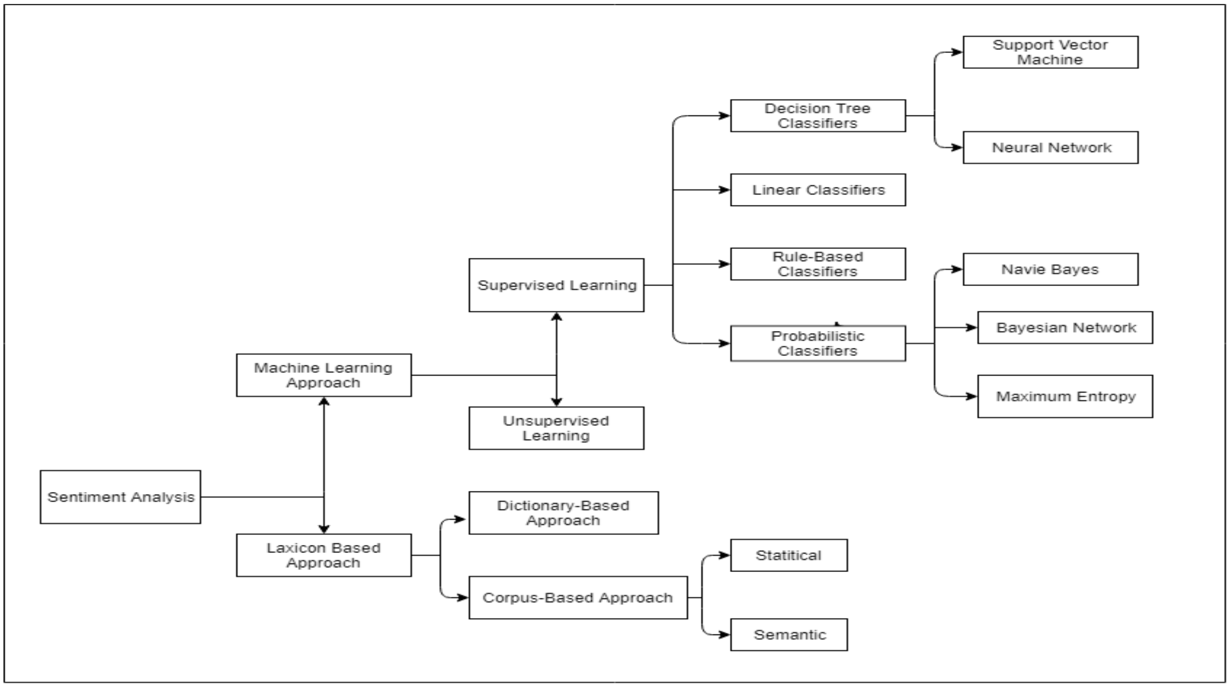
- Product Feature Extraction is a task to extract the product features from its review.

- Opinion Retrieval is a task to searching a specific opinion and gathers it.

3.1.2 Industrial Prospective of Sentimental Analysis

Sentiment analysis techniques are classified into two categories namely lexicon based approach and machine learning based approach.

Lexicon based approach is further divided into two category namely dictionary based and corpus based approach. In dictionary based approach, sentiment is identified using synonym and antonym from lexical dictionary like WordNet, WordStat etc. In corpus based approach, it identifies opinion words by considering word list. Corpus based approach further more classified as statistical and semantic approach. In statistical approach, co-occurrences of words are calculated to identify sentiment. In semantic approach, terms are represented in semantic space to discover relation between terms .



3.1.2.1 Machine Learning Approach

The text classification methods using Machine learning are divided into Supervised and Unsupervised learning methods.

- The supervised learning methods use a large no of training dataset.
- The unsupervised learning methods are used when it's difficult to find in training dataset.

3.1.2.1.1 Supervised Learning

The supervised learning is dependent on existence of previous lebelled training dataset. There are many kinds of supervised classifiers. We will discuss about them in brief the next subsection.

3.1.2.1.1.1 Decision Trees

Decision tree is a very useful technique to get probability of making a correct decision, most of the time. As a method it allows us to approach the problem in a structured and systematic way to arrive at a logical conclusion. It is a decision support tool that uses a tree-like graph or moedel of decisions and their possible consequences, including chance-event outcomes, resource costs and utility.

3.1.2.1.1.2 Probabilistic Classifiers

Probabilistic classifiers are known as the most popular classifier in machine learning. Probabilistic classifiers use mixture models for classification which assumes that each class is a component of mixture and each mixture is a generative model that provides the probability of sampling a particular term for that component. These are called generative classifiers also. Some of the most famous probabilistic classifiers are discussed below.

3.1.2.1.1.2.1 Naive-Bayes Classifiers

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong(naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of Parameters linear in the number of variables (features/predictors) in a learning problem.

It is based on the application of the Baye's rule given by the following formula:

$$P(C=c|D=d)=\frac{P(D=d|C=c)P(C=c)}{P(D=d)}$$

And the simplified Baye's formula can be written as:

$$P(c|d)=\frac{P(d|c)P(c)}{P(d)}$$

c -Hypothesis, d -Tuples, P(c|d) represents Posterior probability of c conditioned on d i.e. the Probability that a Hypothesis holds true given the value of d , P(c) represents Prior probability of c i.e the Probability that c holds true irrespective of the tuple values, P(d|c) represents posterior Probability of d conditioned on c i.e. the Probability that d will have certain values for a given Hypothesis, P(d) represents Prior probability.

3.1.2.1.1.2.2 Maximum Entropy Classifiers

The Maximum Entropy Classifier is known as conditional exponential classifier. It converts labelled feature sets to vectors using encoding process. This encoded vector is then used to calculate weigths for each feature set. This

Classifier is parameterized by a set of weights, which is used to combine the joint features that are generated from a feature-set by an encoding. The encoding maps each feature, label pair to a vector. The probability of each label is then computed using the following equation :

$$P(fs|label)=\frac{dotprod(weights,encode(fs,label))}{sum(dotprod(weights,encode(fs,label))forlinlabels)}$$

3.1.2.1.1.3 Linear Classifiers

Linear classifiers make classification decision based on the value of a linear combination of the characteristics of an object to identify which class or group it belongs to.

Given $\overline{X} = \{ x_1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x_n \}$ is the normalized document word frequency, vector $\overline{A} = \{ a_1 \ldots\ldots\ldots\ldots\ldots a_n \}$ is a vector of linear co-efficients with the same dimensionality as the fiture space , and b is a scalar; the output of the linear predictor is define as $p = \overline{X}\,\overline{A}+b$, which isthe output of the linear classifiers; among them is Support Vector Machine (SVM) which is a form of classifiers that attempt to determine good linear separators between different classes.

3.1.2.1.1.3.1 Support Vector Machine (SVM)

SVM is binary classification algorithm. In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Given a set of points of 2 types in N dimensional place, SVM generates a (N—1) dimensional hyperlane to separate those points into 2 groups.

3.1.2.1.1.3.2 Neural Network (NN)

Neural Network is inspired by biological neural networks in the brain. The neural network Systems can learn to do tasks by being trained on examples, rather than task-specific programming Comprised of layers of interconnected nodes that get activated by an activation function and connections are dealt with by the propagation function. Multiple neural networks are used for non-linear boundaries. Multilayer of Neural network are used to induce multiple piecewise linear boundaries. In neural network, the feed of a layer neuron is the output of its previous network.

The neural network training is a complex process. The NN can be used for text classification and perception learning.

3.1.2.1.2 Unsupervised Learning

Unsupervised learning is useful in cases where the challenge is to discover implicit relationships in a given *unlabeled* dataset (items are not pre-assigned). Unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data. It s a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. In unsupervised methods ; the data set are split into small packets (keys) and categorized each sentence using keyword list of each category and sentence similarity measure.

Unsupervised learning problems can be further grouped into clustering and association problems.

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data.Clustering is the task of grouping a set of objects such that objects in the same group (*cluster*) are more similar to each other than to those in other groups.

- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Some popular examples of unsupervised learning algorithms are:

- **k-means for clustering :** K-Means clustering is an unsupervised learning algorithm that finds a fixed number (*k*) of clusters in a set of data. A cluster is defined by a centroid, which is a point (either imaginary or real) at the center of a cluster. Every point in a data set is part of the cluster whose centroid is most closely located. K-means is an iterative clustering algorithm that aims to find local maxima in each iteration.

Algorithmic steps for k-means clustering

Let X = $\{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Randomly select '*c*' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1 / c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

● **Apriori algorithm for association rule learning problems :** Apriori algorithm proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.Apriori uses breadth-first search and a hash-tree structure to count candidate item sets efficiently.

3.1.2.2 Lexicon Based Approach

Semantic orientation (SO) is a measure of subjectivity and opinion in text. It usually captures an evaluative factor (positive or negative) and potency or strength (degree to which the word, phrase, sentence, or document in question is positive or negative) towards a subject topic, person, or idea .Application of a lexicon is one of the two main approaches to sentiment analysis and it involves calculating the sentiment from the semantic orientation of word or phrases that occur in a text . With this approach a dictionary of positive and negative words is required, with a positive or negative sentiment value assigned to each of the words. Semantic orientation of phrases is determined as positive if it is more related to “best” and is considered to negative if it is more it’s related to “poor”. It is based on opinion lexicon.

- The dictionary based approach which depends on finding option seed words and search dictionary of their synonyms and antonyms.
- The corpus-based approach starts with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations.

3.1.2.2.1 Dictionary-Based Approach

In dictionary based approach, a small set of opinion words are collected. It is English database dictionary where every term is associated with each other via link. Mostly WordNet is used to check similarity with words and to calculate sentiment score. It links to sets of syntactic category which are verb, adjective, adverb and noun.WordNet and Dictionary based approach, both are improved and add new entries (newly found word) after each iteration. It is linked with semantic relations those are termed as synonym, antonym, hyponymy, metonymy, troponymy, Entailment etc .

The major disadvantage of dictionary based approach is inability to fine opinion words in domain and context specific orientation. It is used for ten texting advertisement and in improves and relativeness of user experience.

3.1.2.2.2 Corpus-Based Approach

Corpus linguistics is the study of language as expressed in corpora (samples) of "real world" text. The text-corpus method is a digestive approach that derives a set of abstract rules that govern a natural language from texts in that language, and explores how that language relates to other languages. Originally derived manually, corpora now are automatically derived from source texts. Corpus linguistics proposes that reliable language analysis is more feasible with corpora collected in the field in its natural context and with minimal experimental-interference. Corpus linguistics has generated a number of research methods, which attempt to trace a path from data to theory.

3.1.3 Vader Lexicon

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains. The VADER lexicon is an empirically validated by multiple independent human judges, VADER incorporates a "gold-standard" sentiment lexicon that is especially attuned to microblog-like contexts.We have used it to calculate the sentiment score of words. In this approach, each of the words in the lexicon is rated as to whether it is positive or negative, and in many cases, how positive or negative.

3.2 Conclusion

There are many techniques for sentiment analysis : machine learning and lexicon based. In machine learning approach there supervised and unsupervised method. We have used k-means clustering under unsupervised method which will be discussed next.

CHAPTER 4:

4.1 Proposed Method

4.1.1 Introduction

For the sentiment analysis using twitter data we used machine learning aproach. We have done an API based sentiment analysis for analyzing the sentiment of the tweets. Twitter data are collected and given as input. After that we applied k-means clustering under unsupervised approach.

4.1.2 Proposed Method

The proposed procedure is divided into following steps :

4.1.2.1 Tweeter Data Collection

We have applied some data mining process to collect the twitter data. We have created our own application with help of twitter API. We have collected a large no. of dataset (10000 tweets). For this we have to create a developer account and register our application. Here we received a consumer key and a consumer secret: these are used in application settings and from the configuration page of application we also require an access token and an access token secret which provide the application access to twitter on behalf of the account. The process is divided into two sub-processes. These are discussed in next sub section.

4.1.2.1.1 Accessing the Twitter Data

To make the application and to interact with twitter services we used Twitter provided TWEEPY. We used a bunch of Python-based clients .The API variable is now our entry point for most of the operations we can perform with Twitter. The API provides features to access different types of data. In this way we collected tweets (and more) and stored. The data is stored in JSON format.

4.1.2.2 Data Pre-Processing and Cleaning

In this step we perform the necessary data pre processing and cleaning on the collected dataset. On the previously collected dataset, there are some key attributes (it contains):

- text: the text of the tweet itself
- created_at: the date of creation
- favorite_count, retweet_count: the number of favourites and retweets
- favourite, retweeted: Boolean stating whether the authenticated user (you) have favourite or retweeted this tweet
- Lang: acronym for the language (e.g. “en” for English)
- id: the tweet identifier
- place, coordinates, geo: geo-location information if available
- user: the author’s full profile
- entities: list of entities like URLs, @-mentions, hash tags and symbols
- in_reply_to_user_id: user identifier if the tweet is a reply to a specific user
- in_reply_to_status_id: status identifier id the tweet is a reply to a specific status

We have applied an extensive set of pre-processing steps to decrease the size of the feature set to make it suitable for learning algorithms. The cleaning method is based on dictionary methods.

There are some more steps to process the dataset. They are described below :

4.1.2.2.1 Extracting the tweet text

First we extracted the text from the tweets. As the tweitter data was stored in json format, it is easy to extract the text portion from the tweets by using the json package available in python. Here the other parts of the tweets such as date, tweet id etc. are ignored and we got the text portion from the tweets for the sentiment analysis.

4.1.2.2.2 HTML Character Escaping

Data obtained from twitter usually contains a lot of html entities like &lt; &gt; &amp; which gets embedded in the original data. It is thus necessary to get rid of these entities. To remove them we have used specific regular expressions. We apllied the regex package available in python.

4.1.2.2.3 Decoding Data

This is the process of transforming information from complex symbols to simple and easier to understand characters. The collected data uses different forms of decoding like “Latin”, “UTF8” etc. Foe this, we are change all of this to “UTF-8” for better understanding.

4.1.2.2.4 Stop Word Removal

Stop words are generally thought to be a “single set of words”. We would not want these words taking up space in our database. For this using NLTK and using a “Stop Word Dictionary” we removed the stop words as they are not useful.

4.1.2.2.5 Removal of Punctuations

The puctuations should be dealt with according to the priorities. For example: “.”, “,”,”?” are important punctuations that should be retained while others need to be removed. We replaced every word boundary by a list of relevant punctuations present at that point. We also removed single quotes if exist in the text.

4.1.2.2.6 Other Words and symbols

There are also some information as Hash tags, @, Username, retweet and other modified tweets. All of these are ignored and removed from the dataset. The meaningless words are also ignored by using enchant dictionary package available in python.

4.1.2.3 Calculating the sentimrnt scores

After processing the data we calculated the sentiment score of the words in the processed text. We have applied vader lexicon approach to calculate the sentiment score of the words. It is a very useful tool and gives a decent accurate result. It also calculate scores of empticons present in the tweets.

4.1.2.4 K-means clustering

There are various methods of clustering. K-means is one of the most efficient methods for clustering. In k-mean clustering algorithm, probability of the most relevant function is calculated and using Euclidian distance formula the functions are clustered. The Euclidian distance formula is as follows :

$$D(i,j)=\sqrt{\ } ((X_{i1}-X_{j1})^2 + (X_{i2}-X_{j2})^2 + ... + (X_{ip}-X_{jp})^2)$$

Where :

$D(i,j)$ = distance of i-th data to cluster center j

X_{ik} = i-th on the k-th data attribute

X_{ij} = j-th center point on the k-th data attribute

The letter “k” in the K-means algorithm refers to the number of groups we want to assign in the given dataset. If “n” objects have to be grouped into “k” clusters, k cluster centers have to be initialized. Each object is then assigned to its closest cluster center and the center of the cluster is updated until the state of no change in each cluster center is reached.

We have applied k-means clustering and the steps are as follows :

STEPS :

- Collect the tweets by tweepy
- Store in a file in json format
- Extract the tweet-text from the json file
- Remove URLs, hash tags, symbols like @,$,&,* etc.
- Store the processed text in a text file
- Load the text file and calculate sentiment score by vader-lexicon in the range -1 and 1
- Store the scores in an array and apply k-means clustering

4.1.2.5 Result and analysis

We have calculated the sentiment score of words and emoticons.

Emoticons	Counting	Emoticons	Counting
:)	0.5	‘:]	-0.375
:(	-0.475	-_-	0.365
%)	-0.1	=-s	0.2825
;)	0.325	;(	-0.275
('-:	-0.55	:,’[	0.25
(-*	0.275	:((	-0.25
(;<	0.075	:-/	0.2425
:P	0.725	=]	0.225
(:	0.46	<3=>	0.025
:|	-0.45	:”D	0.205
:>	0.525	=D	0.205
]:D	0.4225	DX	0.2

Fig-1 : Counting the emoticons

After applying k-means clustering we have got 3 clusters (negative, positive and neutral).

Output :

```
[[-0.35625   ]
 [ 0.70178571]
 [ 0.33409091]]
```

K-means clustering is used to predict the sentiment of the text. Our perfomance on grouping of sentiments of social network may be diverted at some point due to presence of spelling mistakes, some words used as both negative and positive sense, use of acronym etc. The results of this work serve as a partial view of the phenomenon. More research needs to be done in order to validate or invalidate these findings, using larger samples.

4.2 Conclusion

Here the procedure of collecting tweets, processing the text, extracting the text portion from the tweets and then applying k-means clustering on the text for sentiment analysis is disscussed. We got the result as 3 clustres i.e. positive, negative and neutral which help us to comment on the sentiment of the tweets.

CHAPTER 5:

5.1 Conclusion and future work

5.1.1 Conclusion

Twitter is a demandable micro blogging service which has been built to discover what is happening at any moment of time and anywhere in the world. In the survey, we found that social media related features can be used to predict the sentiment (positive, negative or neutral) of the tweets.

We have used K-means clustering in the grouping of sentiments of social network. So, our proposed system concludes the sentiments of tweets which are extracted from twitter and grouped during this analysis and we tried to give a prediction on the sentiment of the text.

5.1.2 Future Work

The sentiment analysis problem can be solved to a satisfactory label by manual training. But a fully automated system for sentiment analysis which needs to manual intervention has not been introduced yet. This is one of the main challenges in this field.

One more feature is worth exploring that is whether the information about relative position of word in a tweet has any effect on the performance. As sometime the position of the emoticons and negation words can change the full meaning and polarity of sentences. Besides being the three clusters (positive, negative and neutral) are not equal. So it affects the overall polarity. In this research we are focusing on general sentiment analysis. We can attempt to model few more co-efficient and parameter and degrees in our system, which gives us better accuracy.The better accuracy can be obtained with Deep learning , an emerging and growing field of research in Intelligent Systems

There are several directions to work on increasing the classification accuracy of proposed method. Advanced feature engineering techniques might have significant impact on classifier effectiveness. Moreover, it would benefit of more extensive application of natural language techniques, including part-of-speech (POS) tagging, named entity recognition, abbreviation resolution, relation extraction, etc. Aspectbased sentiment analysis is also one of the fields, which could make use of proposed techniques. Besides , now a days , with the advent of Facebook, Instagram people are expressing their thoughts with pictures and videos along with text. Work on this area can be considered.

5.2 Conclusion

The result obtained from applying k-means clustering is discussed here. The ideas for future work on sentiment analysis is also disscussed.

REFERENCE

Bouras, C. & Tsogkas, V. 2010. "W-kmeans: Clustering News Articles Using WordNet". Retrieved September 5, 2013

Hao Wang, Jorge A. Castanon,Silicon Valley Laboratory,IBM,San Jose, USA, “Sentiment analysis via emoticons using twitter data” 2015 IEEE International Conference on Big Data (Big Data).

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, “Sentiment Analysis of Twitter Data” Department of omputer Science, Columbia University, New York, 2009.

Akshi Kumar and Teeja Mary Sebastian, “Sentiment Analysis on Twitter” department of Computer Engineering, Delhi Technological University, Delhi, India, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.

SanjanaWoonna and PriyankaGiri, “Sentiment analysis of twitter data” International Journal of Innovation and Technology, 2016.

Maks Isa, Vossen Piek. “A lexicon model for deep sentiment analysis and opinion mining applications”. Decis Support Syst 2012; 53:680–8.

B. Pang and L. Lee, "Opinion mining and sentiment analysis" Foundations and trends in information retrieval, vol. 2, pp. 1-135, 2008.

P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V.Stoyanov, and T. Wilson, "Semeval-2013 task 2: Sentiment analysis in twitter" 2013.
Mitali Desai ,Mayuri A. Mehta,Computer Engineering Department,Sarvajanik College of Engineering and Technology,” Techniques for Sentiment Analysis

of Twitter Data: A Comprehensive Survey”,in International Conference on Computing, Communication and Automation (ICCCA2016) .

S. Bhuta, A. Doshi, U. Doshi and M. Narvekar, “A review of techniques for sentiment analysis Of Twitter data”, Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014, pp. 583-591.

V. Singh and S. K. Dubey, “Opinion mining and analysis: A literature review”, in 5th Int. Conf. on Confluence The Next Generation Information Technology Summit (Confluence), 2014, pp. 232-239.

K. Khan, B. Baharudin, A. Khan and F. Malik, “Mining Opinion from Text Documents: A Survey”, Digital Ecosystems and Technologies, 2009, pp. 217-222.

K. Ghag and K. Shah, “Comparative analysis of the techniques for Sentiment Analysis”, in Int. Conf. on Advances in Technology and Engineering, 2013, pp. 1-7.

C.C. Aggarwal and T. Abdelzaher, ”Social sensing,” in Managing and mining sensor data, Springer US, 2013, pp. 237-297.

G. Anastasi, M. Antonelli, A. Bechini, S. Brienza, E. D.Andrea, D. De Guglielmo, P. Ducange, B. Lazzerini, F. Marcelloni, and A. Segatori, ”Urban and social sensing for sustainable mobility in smart cities,” in Proc. of the Sustainable Internet and ICT for Sustainability, 2013. IEEE,2013, pp. 1-4.

Pietro Ducange,Michela Fazzolari,” Social sensing and sentiment analysis on Social Media as Useful Information Source”.

J. Read. “Using emoticons to reduce dependency in machine learning techniques for sentiment classification”. In Proceedings of ACL-05, 43nd Meeting of the Association
for Computational Linguistics. Association for Computational Linguistics, 2005.

B. Liu. “Sentiment Analysis and Opinion Mining”. Morgan and Claypool Publishers: Synthesis Lectures on Human Language Technologies, 2012.
W.Medhat et al. walaa Medhat, Ahmed Hassan, Hoda Korashy, “Sentiment analysis algorithm and application:A survey”, Ain Shams Engneering Journal, p 1093-1113,April,2014.

ggarwal Charu C,Zhai Cheng Xiang.”Mining text data”. Springer New York Dordrecht Heidelberg London, LLC’12;2012.

crtes c,Vapnik V. “Support Vector networks” ,presented at machine learning;1995.
Hatzivassiloglou V,Mckeown K. Predicting the semantic orientation of adjectives. In :Proceedings of ACL’97;1997.

Balakrishnan Gokulakrishnan “Opinion Mining and Sentiment Analysis on a Twitter Data Stream”The International Conference on Advances in ICT for Emerging Regions - ICTer2012 : 182-188.

AizhanBizhanova, Osamu Uchida, “Product ReputationTrend Extraction from Twitter”Social Networking, 2014, 3, 196-202.

Dipak Gaikar ,Bijith Marakarkandy“Product Sales Prediction Based on Sentiment Analysis Using Twitter Data” International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015.

R. Parikh and M. Movassate, “Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009

Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social Media Data", IEEE 14th International Conference on Mobile Data Management ,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, http://doi.ieeecomputersociety.org/10.1109/MDM.2013.

R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

Payne, Thomas E. 1997. Describing Morphosyntax. Cambridge University Press, Cambridge, UK.

van derWouden, Ton. 1997. Negative Contexts: Collocation, Polarity, and Multiple Negation. Routledge, London.

Tottie, Gunnel. 1991. Negation in English Speech and Writing: A Study in Variation. Academic Press, New York.

B. Pang, L. Lee and S. Vaithyanathan, “Thumbs Up? Sentiment Classification Using Machine Learning Techniques,” Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02, vol. 10, pp. 79-86, 2002.

B. Supriyono, “WEB DATA MINING FOR CUSTOMER’S SENTIMENT CLASSIFICATION FOR TELKOM SPEEDY USING TWITTER IN INDONESIAN,” no. August 2015.

Franky, O. Bojar and K. Veselovská, “Resources for Indonesian Sentiment Analysis,” *The Prague Bulletin of Mathematical Linguistics,* vol. 103, no. 1, pp. 21-41, 2015.

S. Das and M. Chen, “Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web,”*Management science, vol. 53 (9), 2004.*

Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

Ali Shah, N. & M. ElBahesh, E. “Topic-Based Clustering of News Articles”, University of Alabama at Birmingham. Retrieved September 23, 2013

D. Manning, C., Raghavan, P. & Schütze, H. 2008. “Introduction to Information Retrieval”, Cambridge, England: Cambridge University Press. Retrieved September 4, 2013

Huang, C., Simon, P., Hsieh, S. & Prevot, L. 2007. “Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Word break Identification”. Retrieved September 2, 2013

Kobayashi, M. & Takeda, K. 2000. "Information retrieval on the web". Retrieved September 20, 2013

Plisson J., Lavrac N. & Mladenic D. 2004. “A Rule based Approach to Word Lemmatization”. Retrieved September 25, 2013

K.Rajalakshmi,, Dr.S.S.Dhenakaran,N.Roobin “Comparative Analysis of K-Means Algorithm in Disease Prediction”, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015

Daljit Kaur and Kiran Jyot, “Enhancement in the Performance of K-means Algorithm”, International Journal of Computer Science and Communication Engineering, Volume 2 Issue 1, 2013

Siddheswar Ray and Rose H. Turi, “Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation”, School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia, 1999