



**SENTIMENT ANALYSIS ON TWEETS BY USING SUBSPACE
CLUSTERING WITH CNN MODEL**

A Dissertation

Submitted

By

M.UMA MAHESH

(11616924)

To

Department of Computer science and Engineering

In fulfillment of the requirements for the

Award of the degree of

Masters of Technology in Computer Science & Engineering

Under the Guidance of

(Mr. Robin Prakash Mathur)

Assistant Professor

Department of Computer Science and Engineering

(DEC, 2017)



TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0030

Supervisor Name : Robin Prakash Mathur **UID :** 14597 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Marothu Uma Mahesh	11616924	2016	K1637	9888530676

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** _____

PROPOSED TOPIC : Clustering High Dimensional Data

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	7.00
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.80
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.20
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.60
5	Social Applicability: Project work intends to solve a practical problem.	7.20
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.00

PAC Committee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): Yes
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): Yes
PAC Member 6 Name: Dr.Priyanka Chawla	UID: 22046	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: Clustering High Dimensional Data

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal

Approval Date: 04 Nov 2017

12/4/2017 12:06:22 PM

Certificate

This is to certify that **M.Uma Mahesh** has completed M.Tech dissertation entitled “SENTIMENT ANALYSIS ON TWEETS BY USING SUBSPACE CLUSTERING WITH CNN MODEL” under my guidance and supervision. To the best of my knowledge, the present work is the result of him original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma. The dissertation proposal is fit for the submission and partial fulfillment of the conditions for the award of M.Tech Computer Science & Engineering.

Date: _____

Signature of Advisor

Name: Robin Prakash Mathur

Uid: 14597

Declaration

I hereby declare that the dissertation proposal entitled, **SENTIMENT ANALYSIS ON TWEETS BY USING SUBSPACE CLUSTERING WITH CNN MODEL** submitted for the completion of M. Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: _____

Investigator:

Regn No:11616924

Abstract

With the advancement in the pervasive technology, there is a spontaneous rise in the size of the data. Such data are generated from various forms of resources right from individual to organization level. Due to the characteristics of unstructured or semi-structuredness in data representation, the existing data analytics approaches are not directly applicable which leads to curse of dimensionality problem. Hence, we are presenting a hybrid model for high dimensional data. Computational analysis is to find the sentiments like positive, negative or neutral of the tweets which are generated by users. We prefer automated sentiment analysis because we have to analyze tweets from large data with high accuracy. By using sentiment analysis, we analyze the sentiments. This proposed methodology formulates the problem of automated sentiment analysis using the concept of deep learning by convolution neural network (CNN). The proposed approach uses Sub-space clustering method and train the features in the convolution neural network. Sub space clustering reduces the overlapping between features of different Classes of sentiment, therefore reduce the false positive rate and increase accuracy of classification.

Acknowledgement

First and foremost, I would like to thank to my mentor of this dissertation, **Asst. Prof. Robin Prakash Mathur** for the valuable guidance and advice. He inspired me greatly to work in this field. His willingness to motivate me contributed tremendously. I would also like to thanks him for showing us some example that related to my field. Besides, I would like to thank the Department of Computer Science of Lovely Professional University for providing me with a good environment and facilities to complete Dissertation task. Finally, an honorable mention goes to my family and friends for their understandings and support.

M.Uma Mahesh

11616924

Table of Contents

PAC Form.....	ii
Certificate by Advisor.....	iii
Declaration.....	iv
Abstract.....	v
Acknowledgement.....	vi
Table of Content.....	vii-viii
List of Figures.....	ix
Chapter 1 Introduction.....	1-8
1.1 Introduction to sentiment analysis.....	1-2
1.2 Tweets	2-4
1.2.1 Sentiment analysis on tweets.....	4-5
1.2.2 Methods to Analyze Sentiments of Tweets	6
1.2.3 Classifier used to classify the tweet during analysis.....	6
1.2.4 Types of Sentiment Analysis.....	7
1.3 Evaluation of sentiment Analysis.....	7-8
1.4 Advantages of Analysis of tweets.....	8
1.5 Disadvantage of Analysis of tweets.....	8
Chapter 2 Literature Review.....	9-14
Chapter 3 Scope of study.....	15
Chapter 4 Objective of the study.....	16
Chapter 5 Research Methodology.....	17-18
Chapter 6 Expected outcomes.....	19
Chapter 7 summary and conclusion.....	20
References	21-23

List of Figures

Figure 1.1 Tweet.....	3
Figure 1.2 Form of Tweets.....	4
Figure1.3 process of sentiment analysis.....	5

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO SENTIMENT ANALYSIS:

First of all, we are going to give a highlight on sentiments. Sentiments are the attitude, emotion or feeling towards a particular thing or subject. The sentiments are nothing but a thought or opinion of the peoples, sentiment can be termed as a simple review of a subject or an object. The emotional significance of a passage or expression as distinguished from its verbal context is termed as sentiments [8]. Now we discuss the sentiment analysis, it is a task of finding the opinions and attitude of people towards specific topics of interest. Be it is a product or a movie, opinions of people matter, and it affects the decision-making process of people.

The first thing a person does when he or she wants to buy a product online is to see the kind of reviews and opinions that people have written. Social media such as Facebook, blogs, Twitter has become a place where people post their opinions on certain topics. The sentiment of the tweets of a particular subject has multiple usages, including stock market analysis of a company, movie reviews, in psychology to analyze the mood of people that have a variety of applications, and so on. Sentiments of tweets can be categorized into many categories like positive, negative, neutral, extremely positive, extremely negative, and so on. The two types of sentiments considered in this classification experiment are positive and negative sentiments [3].

Sentiment analysis is a process of extracting or to identify the sentiment content of text unit by using various methods like NLP, machine learning etc. It is a computational analysis to find the sentiments like positive negative or neutral of the material that is generated by users. We prefer automatic sentiment analysis because we have to analyze material from large data with high accuracy. By using subjectivity analysis, we analyze the sentiments. In this analysis, we determine the subjectivity or objectivity of the text data. In the second step, we check the sentiments and its quantity in the data. On the basis of text and images, we used to analyze the sentiments from given data.

The main goal of sentiment analyses is to evaluate the state of mind of the speaker. This analysis is used to differentiate the opinion of users or speaker on the basis of its binary polarity. There are three primary levels of sentiment analysis a) document level, b) sentence level, and d) aspect-

level. The document-level analysis is used to characterize the feelings of the document. The sentence-level analysis is used to indicate the notion in a sentence. The aspect-level analysis is used to sort the conclusion with respect to the specific aspects of the elements.

1.2 TWEETS

A tweet is a small group of 280 characters or less, which is in text, image and in video form. Tweets are posted on twitter which shows the people interest or opinion related to issues (social or political) and products advertised in the market. People post tweets on Twitter to show the attention, promotion, and joy. These tweets provide an instant update from friends, experts, family and news on the day to day events [2].

Tweets are the textual form of Twitter where we use to communicate or share our activities. Tweets are a collection of textual data and report everything from the daily stress of life to the latest local and worldwide events. The contents show the real-time events in day to day life or daily routine, these contents are full of social information and temporal attributes. This information, data used to analyze sentiments of human beings, because every person uses to express their view on social sites. After analysis of these data valuable information can be extracted that helps to predict any situation or result easily. Twitter gives fine-grained information about each and every-events, instances, perspectives etc. [2]

Following are the types of tweets:

- **Regular tweets:** It is in textual form.
- **Image tweets:** It is the form of a tweet in an image with textual information.
- **Video tweet:** It is a form of video message having the duration of 30 seconds only.
- **Media-rich link tweets:** It is a form of web link it may be an image or video on the website. It is used as a reference link.

Twitter is a social site or a platform where tweets are posted in its different form. In twitter 320 million users active monthly. On Twitter, we make our follower and follow them back. We can share the activities we are doing and make the private account. If we want to text anyone privately we can text easily with complete security [3]. In short, Twitter is a social site where user posts and interacts with the different person connected with Twitter. It is important to go through by registration to see the post and to interact with particulars. If registration is not done, then that user can only read the blog, tweeted by other users they wouldn't be able to post any

tweets by themselves. Users can access through their website. It has more than 25 offices around the world. We can say that all the communication or posting whatever is written on the twitter is the form of Tweets whether it was in a textual form of an imaginary form. In the chapter, we are going to study the sentiments of tweets. The sentiments are the emotion or feeling of the user which they express in the form of tweets on social networking site Twitter [3].



Fig 1.1: Tweet [19]

There are various types of sentiments of a user, like negative if he was disappointed with any services or dis-like any service or positive if he like or satisfied with any service or neutral. Here we are going to analysis the sentiments of tweets from Twitter [5]. There is an example of the form of a tweet, which is used to express their views, whether it was positive or negative or wants to express their day-to-day activity. People use tweets to express their feelings on any topic or activity, whether it was social or political or it may be included in their daily routine. Tweets are formed with the name of which a person used to talk with each other and expresses its feelings.

In the analysis of Twitter data, there is a major issue is faced that is to separate mixed and noisy data information from valuable of needed data in real-world events. There are many scalable and significant approaches required for handling and processing a large amount of Twitter data. Another difficult issue in extracting sentiments is to understand the meaning of words. The main

reason for this issue is the length of tweet message, formal, informal, regular, irregular words and spelling and grammatical errors [3].






<p>1: An image file (only text information) is used for disseminating alert.</p>	<p>2: An image is used for emphasizing significant information of the alert.</p>	<p>3: An image is used for advertising a product.</p>
<p> Ricardo Ordieres @RicardoOrdieres</p> <p>Very unfortunate...I wish I could be there to comfort anyone in need. Boston please stay safe. Pass this around! pic.twitter.com/5TL3CewLyz</p> <p><small>tarzanebingoffemales:</small> Do NOT drive through Boston or take the subway right now. The emergency radio is buzzing with more possible bombs. Please spread th everywhere you can to let your friends and family members know. It coul save a life.</p> <p>1:25 PM - 15 Apr 2013</p>	<p> Boston Police Dept. @Boston_Police</p> <p>#WANTED: Police seeking MA Plate: 116-GC7, '99 Honda Sedan, Color - Green. Possible suspect car. Do not approach. pic.twitter.com/IVCPtmVwRTb</p>  <p>11:09 AM - 19 Apr 2013</p>	<p> Saucony @saucony</p> <p>Now available for pre-order, the #BostonStrong Lace Medallion. 100% of every \$5 pair sold goes to The One Fund Boston pic.twitter.com/IB2AlwgHqw</p>  <p>7:15 PM - 25 Apr 2013</p>

Fig 1.2: Form of tweets [20]

The above image is giving the description of the different form of tweets. As we are going to analyze the sentiments of tweets, firstly we have the knowledge of sentiments what it is, how it works and what is the need of that so there is a brief description of sentiment analysis [20].

1.2.1 Sentiment Analysis of Tweets :

Sentiment analysis is a process of extracting or to identify the sentiment content of text unit by using various methods like NLP, machine learning etc. It is a computational analysis to find the sentiments like positive negative or neutral of the material that is generated by users. We prefer automatic sentiment analysis because we have to analyze material from large data with high accuracy. By using subjectivity analysis, we analyze the sentiments. In this analysis, we determine the subjectivity or objectivity of the text data. In the second step, we check the sentiments and its quantity in the data. On the basis of text and images, we used to analyze the sentiments from given data.

The main goal of sentiment analyses is to evaluate the state of mind of the speaker. This analysis is used to differentiate the opinion of users or speaker on the basis of its binary polarity. There are three primary levels of sentiment analysis a) document level, b) sentence level, and d) aspect-level. The document-level analysis is used to characterize the feelings of the document. The sentence-level analysis is used to indicate the notion in a sentence. The aspect-level analysis is used to sort the conclusion with respect to the specific aspects of the element analyze the sentiment where the combination of different emotions is included. We resolve such kind of problem with the classification process on different levels of text like terms, phrase, sentence or document. In sentiment analysis, most of the work has been done to find sentiment regarding a general topic by taking an assumption that viewer talks about an individual topic. In such reviews, it is easy to analyze the sentiments of the subject [6].

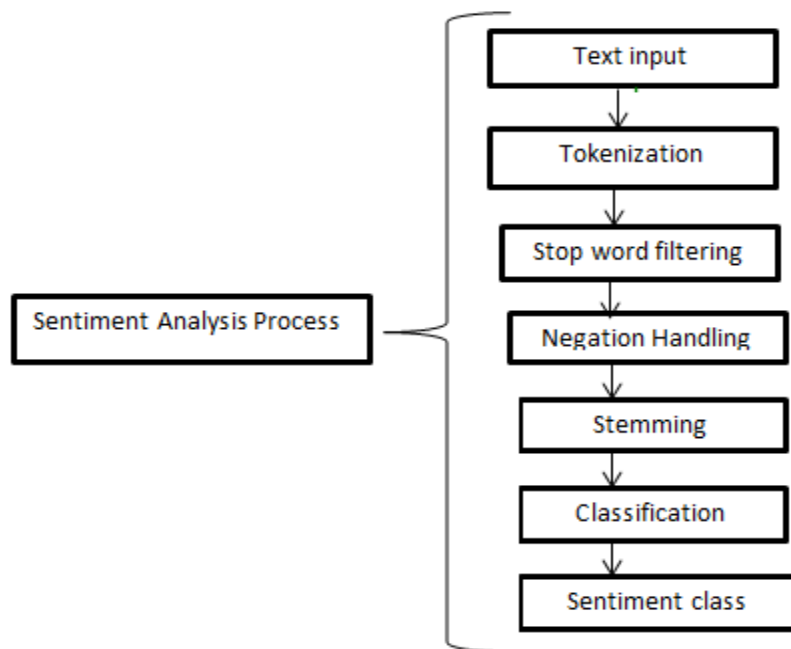


Figure 1.3 Process of sentiment analysis

1.2.2 Methods used to analyze the sentiments of tweets

There is three type of method used to analyze the tweets:

a) Knowledge-based Technique: As name resembles, in this technique we collect the knowledge about the subject and the strength of the sentiments. As we discuss there is various type of sentiments in textual analysis like sad, happy, good, and sorry. In this technique, we collect all the data from different sources like television, internet, social sites etc.

b) Statistical Method: In this technique, we use such elements like Latent Semantic Analysis (LSA), SVM and point wise mutual Information (PMI). It is most sophisticated methods to detect the sentiments properly.

c) Hybrid Approaches: it works on both machine learning and elements from collected data. The detection is done through the analysis of the concept that doesn't explicitly convey relevant information, but which are implicitly linked to another concept.

1.2.3 There are various classifiers used to classify the tweet during analysis:

- a) Hybrid classifier
- b) Mix based classifier
- c) SVM
- d) MAX entropy

Classifiers are used for the purpose of automatically labeling tweet sentiments by using emoticons. By using unigrams and bi-grams features it gives the accuracy of 80% on 0.5 labeled tweets. For the purpose of sentiment classifications semantic and synthetic approaches are commonly used. Unsupervised, supervised and semi-supervised as a summary we use these approaches while on the basis of previous work semantic, synthetic, link-based and Stylistics based approaches are used for the classification of sentiments in tweets [3].

1.2.4 Types of sentiment analysis:

1) Manual Processing

Most accurate and mature judge of sentiments is human interpretation but it does not give 100% accuracy. Now a day this process is rarely used without any additional tool. It always needed a substitute tool because of growth of social media.

2) Keyword processing

This processing works on single individual words having the degree of polarity positive or negative, then after it gives an overall result in percentage. Positive words consist of love, happy, smile, yippee etc. and negative words like anger, aggression, dislike etc.

The main merit of this process is it is fast in processing; cheap in cost and predictable while implementation.

There are many disadvantages of these processes:

- 1) It cannot deal with polarized word combined.
- 2) It cannot deal with the words having the different meaning.
- 3) It is not able to deal with multiple words.

3) Natural language processing:

This process can also be termed as NLP, data mining, text analytics. It is a computer system that used to process human language in terms of meaning. It is used to understand the several words having the different meaning, whether it was in phrase, sentence or in ideas. It works on analyzing the meaning of words.

1.3 EVALUATION OF SENTIMENT ANALYSIS:

The main objective of sentiment analysis is to analyze the sentiments which match with the human judgment accurately. There are two categories like positive or negative on which our precision and accuracy of analysis majorly depend. There are two types of sentiments as we discuss above are of various type positive or negative or neutral. In sentiments analysis, words float with different emotion some having positive attitude whereas some having negative attitudes hence there is a simple one –dimensional model of sentiments from negative to positive.

It is seen through research that 70% accuracy in classifying sentiments. The computer system will make very different errors than human assessor. The computer system will have the trouble with negation, exaggeration, jokes or sarcasm which is easy to handle for a human reader. Sentiment analysis is a very useful process for any organization as it indicates the emotion of a human being in every individual circumstance.

1.4 Advantages of Analysis of tweets:

- 1) It is an instant way to identify the opinion of users.
- 2) It gives the ability to work as per the ended of consumers or customers.
- 3) It elaborates the SWOT analysis of any organization
- 4) It helps to develop that product what a customer actually wants.
- 5) It increases the accuracy of the productive items.

1.5 Disadvantages of analysis of tweets:

- 1) Noisy text
- 2) Use of generic lexicons in lexeme feature space
- 3) Lexicon is not capable of analysis the emotion like wow, hurray, yippee etc.
- 4) Only 70 % accuracy is given after analysis.

Chapter 2

Literature Review

Lim, et al. [1] a latent infectious disease has come to the public knowledge which is not known by public health department or institutes earlier. About this existing disease, there are some important but unknown approaches are existing like disease type or name data consist of national public health institutes in the social media. While data processing stage there is some important information that is extracted. In this paper, an unsupervised sentiment analysis is proposed to evaluate the information about the disease. Here we able to identify the symptoms, body part pain location can be identifying by using social media data. With the help of extracted data, we can easily create weighing vectors to resolve the disease. We use real electronic medical record for 104 individuals have been gone through diagnosis with influenza in the interval of August 2012 to May 2013. The result demonstrates the highest precision, recall, and f-score value.

Kumar, et al. [2] assessing the precision of various machine learning calculation for the errand of twitter notion investigation. As tweeter is involved with 190 million tweets in a day and also contains a number of sentiments over them about the object, feeling like about anything. AS these tweets are related to our day to day life it is important to utilize these tweets in a helpful manner hence we use two territories for identifying or extracting useful data from tweets. Here we used to assess an examination for a general conclusion from a small blog on twitter.

AL-Sharuee, et al. [3] gives a highlight on the analysis of sentiments. There are two phase 1) Contextual analysis and the second one is unsupervised learning approaches. The first phase consists of 1) data interpretation 2) spelling correction 3) intensifier handling 4) negation handling 5) contrast handling whereas the second phase consists of clustering of classifiers by using majority voting mechanism. This classifier is a modification of K-means algorithm and we used to modify base classifier by using extracting initial centroids from Sent word net (SWN). Here we also discuss the problem with Australians airline regarding sentiment analysis they offer benchmark problem of analysis. The experimental result shows that there is an improvement in clustering performances in terms of accuracy, stability, and generations.

Gokulakrishnan, et al. [4] an advertised stream of tweets from the Twitter micro blogging website is preprocessed and arranged in light of their passionate substance as positive, negative

and superfluous; and examinations the execution of different ordering calculations in view of their accuracy and review in such cases. Further, the paper epitomizes the uses of this examination and its constraints.

Anjaria, et al [5] it's an approach to exploiting the user influence factor. This kind of approach is used to predict the results of an election. By using direct and indirect features of data collected from twitter which can be extracted by using hybrid methods. These twitter data is based on super revised classifiers like SVM, Naive Bayes, and Artificial Neural Networks. For the dimensionality reduction, we use SVM with the combination of principal component analysis (PCA). This paper gives two different case study of different scenario i.e. US presidential election in 2012 and other is Karnataka Assembly Election in 2013. From these two studies, we try to determine the point where the prediction by twitter data fails means where our process fails. And it seems by the experimental method that SVM predicts with only 58% accuracy in case of Karnataka.

Das, et al. [6] explanation of detailed work done while developing the system which will further used for the analysis of sentiments in tweets. The works as extracting data from social media or tweets from tweeter posts, then its pre-processing and then it connected with Alchemy API by REST call method. This method gives or demonstrates the result in the form of the graph. There is an analysis done for collecting sentiments for Samsung Galaxy from the proposed system.

Lei Zhang, et al [7] defined the opinion mining issue. From the definition, the key specialized issues that should be tended to. We at that point portray different key mining assignments that have been examined in the exploration writing and their agent procedures. After the discussion about the issue of recognizing sentiment spam or fake audits, finally, additionally present the exploration point of surveying the utility or nature of online audits.

P., and Mohammad, et al. [8] identified opinionative information in the Web and classifying them according to their polarity, i.e., regardless of whether they convey a positive or negative meaning. Slant Analysis is an issue of content based examination, however there are a few difficulties that make it difficult when contrasted with customary content based investigation This plainly expresses there is need of an endeavor to work towards these issues and it has

opened up a few opportunities for future research for handling nullifications, shrouded estimations identification, slangs, polysemy. Be that as it may, the developing size of information demands programmed information investigation procedures. In this paper, a point by point overview of different strategies utilized as a part of Sentiment Analysis is done to understand the level of work.

Preslav, et al. [9] advancement and assessment of a semantic examination undertaking that lies at the crossing point of two exceptionally popular lines of research in contemporary computational etymology: (1) estimation investigation, and (2) normal dialect preparing of web-based social networking content exist. The errand kept running in 2013 and 2014, pulling in the most astounding number of taking interest groups at Semantic Evaluation in the two years, and there is a progressing version. The errand incorporated the making of a substantial relevant and message-level extremity corpus comprising of tweets, SMS messages, Live Journal messages, and an exceptional test set of mocking tweets. The assessment pulled in 44 groups in 2013 and 46 of every 2014, who utilized an assortment of methodologies. The best groups could outperform a few baselines by sizable edges with change over the 2 years the errand has been run.

Meral, et al. [10] sentiment analysis has been performed by collecting data from the Twitter. To perform this analysis an intelligent system has been created by using machine learning approaches such as Naïve Bayes, Random Forest, SVM and the compared results has been given

Sagar, et al. [11] analyze lexicon approaches and learning based a method which is used to analyze the sentiments from the text. There are many difficulties and problems introduced while analyzing the sentiments of textual data from social sites like Face book, Twitter etc. In this paper, explained issues are highlighted by authors.

Yulan He, et al. [12] in this paper, there is an introduction semantic feature into the training sets for sentiment analysis. Addition of semantic concept as additional features for each removed entity. This procedure helps in measuring the correlation of the concept with negative\positive sentiments. We use the approach to expect sentiments for three informational indexes of twitter.

This will give the outcome which demonstrates a normal increase of harmonic precision by 6.5 % for identifying both negative and positive sentiments and 4.8 % over the gauge of unigrams.

Divakar Yadav, et al. [13] investigate the data there is a discussion on the sentiments analysis for client survey. The investigation is about to determine the polarity of the collected data. For this purpose we first pre-handled the dataset then we used to extract the script from the datasets, at that point select the component vector list and then connect that from machine learning on the basis of arranged calculation in particular. Naive Bayes, Maximum entropy and SVM alongside the Semantic Orientation based Word Net which separates equivalent words and similitude for the substance highlight.

Phienthrakul, et al. [14] in this paper, they broke down and think about the different non-negative linear combination of the kernel. These kernels are connected on product surveys to decide if an audit is sure or negative. The outcomes demonstrate that the execution of the mix bits that outflanks the single kernels.

Efstratios, et al. [15] arrangement of unique philosophy based systems towards a more effective sentiment analysis on tweets. The curiosity of the given approach is that tweets are not just described by a sentiment score, similar to the case with machine learning-based classifiers; however rather, get a sentiment review for the individual particular idea in tweets. Generally speaking, our introduced design brings about a more itemized analysis of post sentiments with respect to a particular point.

Khan, et al. [16] analyzed to detect the urban marvels like people groups' enthusiasm for specific themes or shift of enthusiasm from one subject to another. Some of the time, even only the proportion of tweets identified with a theme showing up in the day by day Twitter can give a decent sign about people groups' level of enthusiasm for that point on that specific day. Unfortunately, the vast majority of the tweets are not expressly labeled with subject keywords by the Twitter clients. Here a technique for programmed labeling of untagged tweets. Their technique depends on identification of important collocations from a huge preparing set of tweets. We could accomplish 88.25% accuracy with high exactness and review.

Chen Min, et.al. [17] Introduction of the background of big data and give a review on related technologies like cloud computing, Hadoop etc. After introduction authors concentrate on 4 phase of big data a) data generation b) data acquisition c) data storage d) data analysis. They introduce general background for every phase and also used to discuss the technical challenges and latest reviews. At last, they examine the application of data which consist of enterprise management, internet of things, online social networks etc.

Ming Hao et al [18] three novel times based visual sentiments analysis techniques are introduced to explored large tweets in this paper. The techniques are a) topic based sentiment analysis that is used to extract maps and measure customer opinion b) stream analysis that is used to identify interesting tweets on the basis of their polarity c) pixel-cell based sentiment analysis that is used to visualize the volume of large data in a single view. These following techniques are used to demonstrate distribution and pattern.

Isah et al. [19] the author analyzing the user's view and experiences related to the drugs and cosmetics by using the concept of machine learning, text mining, and sentiment analysis. Data used for the proposed analysis was Twitter data and Facebook comments. These comments are used to find out the user's view related to the brands. This method is helpful for the companies to improve the quality of the products when user gave negative comments.

Li, Jinyan, et al. [20] evaluated the various classification algorithms with the filtering schemes. These filtering schemes reduce the original dataset with respect to the contextual polarity. This approach called "Hierarchical classification". A different kind of schemes and algorithms are discussed over three sets of news articles. Binary and multi-class classification is applied to this data.

Rao, Yanghui, et al. [21] discussed the problem of social emotion mining of the online users in the news domain. The author proposed two models for modeling topics and emotions. The first model is supervised model which generates a set of topics from words. The second model generates topics from social emotions directly. Both models can generate the social emotion lexicon samples.

Akaichi et al. [22] the author mainly focused on the Text mining for sentiment classification. The classification process is performed on the Tunisian user's Facebook statues. The main aim is to extract the useful information related to the user's behavior during some specific period of time. A method base on Support Vector Machine (SVM) and Naïve Bayes is proposed for the classification.

CHAPTER 3

SCOPE OF THE STUDY

Sentiment analysis is a process of extracting features from user's thoughts, views, feeling and opinion. Analyses of tweets express the latest trends regarding political issues, social issues, products by tweets. Automatic analyses of tweets increase the understanding of stream. In the proposed approach, we increase the pattern feature and cluster the tweets by subspace clustering approach. This reduces the overlapping of effective features and Increase the true positive rate by clustering of tweet labels.

CHAPTER 4

OBJECTIVE OF THE STUDY

1 To analyze the tweets and classify according to its sentiments:

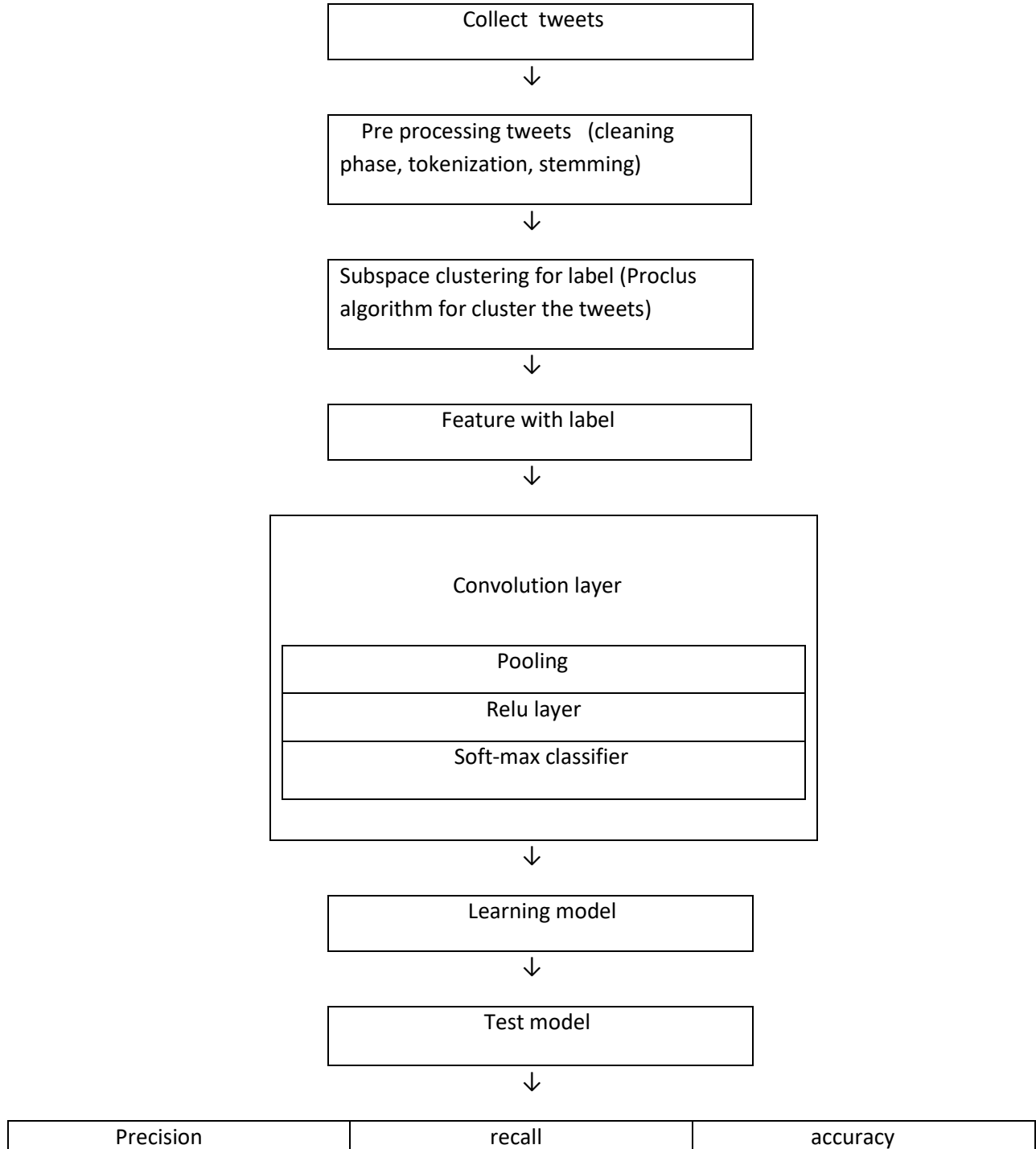
In this objective, we collect the tweets using REST API and analyze the data with static mean, standard deviation by converting the text data into vector form and calculate the number of positive, negative and neutral tweets. This analysis is to understand the data and features of the text.

2 To propose subspace clustering with CNN and evaluate precision, recall, and accuracy.

In this objective, we preprocess the tweets by using tokenization, stop word removal and stemming, then the clustering of tweets according to its subspace labels, and CNN will process the task of learning. Results will be evaluated by using precision which defines the false positive rate and recall which defines the false negative rate of tweets during the prediction process done by using CNN model.

CHAPTER 5

Research Methodology



Methodology process:

In this process, we collect the tweets by using API. Then we use tokenization, stop word filtering for preprocessing of collected tweets. After preprocessing of collected tweets labeling and extraction of features is performed from collected tweets. Then after these extracted features are placed in convolution layer and start the pooling process. Input is given by pooling layer in relu layer and by using soft-max classifier output is given relu layer, output result is given to learning model by the classifier. The output after learning process we test the given output in the test model. At last, we calculate the parameters like precision, recall, and accuracy of the collected tweets.

CHAPTER 6

Expected Outcomes

In automated tweet sentiment analysis process the main challenge is accuracy and multiclass of tweets. In the proposed approach CNN is used for learning which generates different patterns of features by using activation function and then classify these features by using soft-max classifier. Different patterns generated by the activation function increases the domain information of text and then following outcomes will come.

- 1) Reduce the false positive rate of classification which makes an automated system.
- 2) Reduce the overlapping of tweets by subspace clustering which increases the accuracy.

CHAPTER 7

Summary and Conclusion

Sentiment analysis is a process which is used to analyze the thoughts, opinion, and feelings of the person by using textual data. In this research, sentiment analysis is performed on tweets. A tweet is a group of characters which explains the views or opinion of the peoples regarding the issues all around. In this work, tweets are pre-processed and these tweets are clustered for labeling the features. Then we apply CNN to the labeled features for classification. At last, we are going to analyze the results by using following parameters precision, recall, and accuracy.

List of Reference

- [1] Lim, Sunghoon, Conrad S. Tucker, and Soundar Kumara. "An unsupervised machine learning model for discovering latent infectious diseases using social media data." *Journal of biomedical informatics* 66 (2017): 82-94.
- [2] Kumar, Praveen, et al. "Analysis of Various Machine Learning Algorithms for Enhanced Opinion Mining Using Twitter Data Streams." *Micro-Electronics and Telecommunication Engineering (ICMETE), 2016 International Conference on*. IEEE, 2016.
- [3] AL-Sharuee, Murtadha Talib, Fei Liu, and Mahardhika Pratama. "An Automatic Contextual Analysis and Clustering Classifiers Ensemble approach to Sentiment Analysis." *arXiv preprint arXiv:1705.10130* (2017).
- [4] Gokulakrishnan, Balakrishnan, et al. "Opinion mining and sentiment analysis on a twitter data stream." *Advances in ICT for emerging regions (ICTer), 2012 International Conference on*. IEEE, 2012.
- [5] Anjaria, Malhar, and Ram Mohana Reddy Guddeti. "Influence factor based opinion mining of Twitter data using supervised learning." *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on*. IEEE, 2014.
- [6] Das, T. K., D. P. Acharjya, and M. R. Patra. "Opinion mining about a product by analyzing public tweets in Twitter." *Computer Communication and Informatics (ICCCI), 2014 International Conference on*. IEEE, 2014.
- [7] Liu, Bing, and Lei Zhang. "A survey of opinion mining and sentiment analysis" *Mining text data*. Springer US, 2012. 415-463.
- [8] Patil, Harshali P., and Mohammad Atique. "Sentiment analysis for social media: a survey." *Information Science and Security (ICISS), 2015 2nd International Conference on*. IEEE, 2015.
- [9] Nakov, Preslav, et al. "Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts." *Language Resources and Evaluation* 50.1 (2016): 35-65.

- [10] Meral, Meric, and Banu Diri. "Sentiment analysis on Twitter" *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*. IEEE, 2014.
- [11] Bhuta, Sagar, et al. "A review of techniques for sentiment analysis Of Twitter data." *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*. IEEE, 2014.
- [12] Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." *The Semantic Web–ISWC 2012* (2012): 508-524.
- [13] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." *Contemporary computing (IC3), 2014 seventh international conference on*. IEEE, 2014.
- [14] Phienthrakul, Tanasanee, et al. "Sentiment classification with support vector machines and multiple kernel functions." *Neural Information Processing*. Springer Berlin/Heidelberg, 2009.
- [15] Kontopoulos, Efstratios, et al. "Ontology-based sentiment analysis of twitter posts." *Expert systems with applications* 40.10 (2013): 4065-4074.
- [16] Khan, Muhammad Asif Hossain, Masayuki Iwai, and Kaoru Sezaki. "Towards urban phenomenon sensing by automatic tagging of tweets." *Networked Sensing Systems (INSS), 2012 Ninth International Conference on*. IEEE, 2012.
- [17] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." *Mobile Networks and Applications* 19.2 (2014): 171-209.
- [18] Hao, Ming, et al. "Visual sentiment analysis on twitter data streams." *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011.
- [19] Isah, Haruna, Paul Trundle, and Daniel Neagu. "Social media analysis for product safety using text mining and sentiment analysis." *Computational Intelligence (UKCI), 2014 14th UK Workshop on*. IEEE, 2014.
- [20] Li, Jinyan, et al. "Hierarchical classification in text mining for sentiment analysis of online news." *Soft Computing* 20.9 (2016): 3411-3420.
- [21] Rao, Yanghui, et al. "Sentiment topic models for social emotion mining." *Information Sciences* 266 (2014): 90-100.

[22] Akaichi, Jalel, Zeineb Dhouioui, and Maria José López-Huertas Pérez. "Text mining facebook status updates for sentiment classification." *System Theory, Control and Computing (ICSTCC), 2013 17th International Conference*. IEEE, 2013.