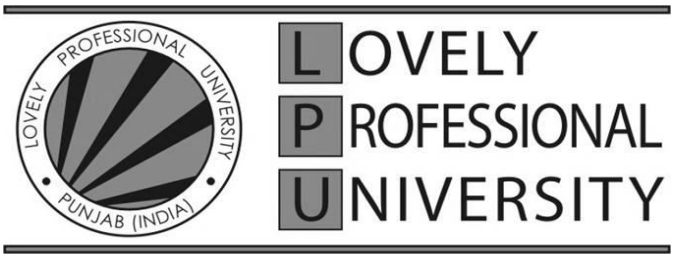

SENTIMENT ANALYSIS ON MOVIE REVIEWS USING HYBRID CLASSIFICATION ALGORITHM

A Dissertation Proposal

Submitted by

Furqan Iqbal

(11607261)

to

Department of Computer Science

In fulfilment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science & Engineering

Under the guidance of

Mr. Chirag Sharma

(NOVEMBER,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 :** CSERGD0322

Supervisor Name : Chirag Sharma **UID :** 16717 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Furqan Iqbal	11607261	2016	K1637	8825024287

SPECIALIZATION AREA : Program Methodology and Design **Supervisor Signature:** _____

PROPOSED TOPIC : sentiment analysis using classification algorithms

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.	7.40
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.	7.00
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.00
5	Social Applicability: Project work intends to solve a practical problem.	7.00
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.20

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): Yes
PAC Member 2 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 3 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): Yes
PAC Member 5 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 6 Name: Harleen Kaur	UID: 14508	Recommended (Y/N): NA
PAC Member 7 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 8 Name: Tejinder Thind	UID: 15312	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): NA

Final Topic Approved by PAC: **sentiment analysis using classification algorithms**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal **Approval Date:** 04 Nov 2017

11/30/2017 1:04:31 PM

DECLARATION

I hereby declare that the proposal for dissertation entitled **Sentiment Analysis On Movie Reviews Using Hybrid Classification Algorithm** submitted for the partial fulfilment of M.Tech Degree is completely my original work having acknowledged all the references and ideas. It does not contain any work for the award of any other degree or diploma.

Date: Furqan Iqbal

11607261

CERTIFICATE

This is to certify that **Mr. Furqan Iqbal** has completed M.Tech dissertation proposal titled **Sentiment Analysis On Movie Reviews Using Hybrid Classification Algorithm** under my guidance and supervision. To the best of my knowledge, the presented work is a result of his 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 the partial fulfilment of the Conditions for the award of M.Tech Computer Science & Engineering.

Date:

Mr. Chirag Sharma

(Advisor)

ACKNOWLEDGEMET

I am very thankful to my mentor **Mr. Chirag Sharma,** Asst. Prof. in Department of Computer Science & Technology, Lovely Professional University, for his patience and advisement in my dissertation work. Without him helping me I would not have been successfully in my work. I would also like to show my gratitude to **Mr. Dalwinder Singh**, Professor and Head, Department of Computer Science and Engineering, Lovely Professional University, for giving students guidance whenever they needed it and keeping the university resources available for the students. I would also like to thank my friends helping me and coming to my aid whenever I needed moral support or help. Finally, I would like to show an immense amount of appreciation towards all the people directly or indirectly contributed to the development and success of this work.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

In today's lifestyle movies are a major source of entertainment for the people and a business worth millions. The online movie reviews play a great part for the people to choose the movie they wish to watch. The analysis of the opinions about a movie can become a deciding factor for the success of a movie. The people can get better understanding about the movie's acceptability by general audience through the sentiment analysis of the movie reviews.

In our research work we have proposed to use Gini Index method for feature selection and using ensemble learning technique for classification using Naive Bayes, Maximum Entropy and Gradient Descent algorithms on a given movie dataset.

CHAPTER 1
INTRODUCTION

INTRODUCTION

In these times of Social media, high speed communication is becoming increasingly popular considering that mobile engagement and high speed internet which has made it possible to share data between anyone, anywhere on the planet. Social media is considered a very versatile topic in many fields of research. As quantity of individuals utilizing social network is increasing on a daily basis to be in contact with their peers in order that they can share their private or non-private messages, a large amount of data is accumulated because of it.

Social Media these days can keep an eye on our whole life like a digital diary. These days many online businesses are being set up that are using data from Social Media. On different social media sites advertisement of many kinds of products are being done to improve the business [2]. To advertise their products the companies must be well informed about the impact of social media coverage because it can be positive as well as negative.

To create a setup for monitoring the data on the social media sites, various tools are required which entails two things: first of all to measure what are number of web users is their brand attracting and second to discover what the people who use it think with regards to the product manufacturer.

To measure the opinion of the customers is not always an easy task. Knowing their perspective cannot be done through a simple math formula and thus we require to perform Sentiment analysis, which is identifies the polarity of user for his purchasing habits, the target subject and the feelings about certain type of things.

Operating processes from this junction, where many of the builders are facing a lot of trouble when they are trying to seek and find their own instruments of sentiment analysis can be a huge task. For more than the past decade, an inquisitiveness has emerged in the minds of people from the information technology fields been about the use of social media evaluation for promoting, opinion evaluation and working out a cohesion group. Online information from many social media sites is having many different kinds of features. Some of the features depend on the size or type of the data. Analyzing the human sentiment used to be something for philosophers and social science students but today it a something that has become a billion dollar industry.

Social media is a humongous platform for people to show the perspective and take a stand whenever it is needed and meet new people in today's fast paced world. Every day people post millions of Tweets, Facebook and videos on the Youtube. The new approaches are coming to light where data is being processed captured at a very fast pace and complex datasets are analysed has happened because of the boost of encouragement that the developers got from the untapped resources of digital world. Various techniques have been made possible by the developers. The different approaches that have been made possible because of big data can give new insights of human behaviour under the light of sentiment analysis. Different companies in the world are using mere human behaviour to sell their products. This data can be used to aid in influencing our behaviour in the future [1].

1.1 Sentiment Analysis

In today's world more data has been collected in the year 2016 than since the beginning of civilization. We now know that 90% data that we have accumulated has been created in the 12 months of 2016 [16]. This humongous amount of data is collected through various sources like healthcare, automobiles, retail stores, businesses, automated sensors, social media, biometric information and information from even human bodies. This amount of data if analyzed can give new insights for the conception and production of new ideas, products or services that can increase productivity, organizational profits and even make human life simpler, better and more secure. The majority of this data is raw and it cannot be directly used for analysis to get a meaningful result. The data without structure might be around 70% to 80% [17]. The natural language processing deals with the interaction between the computers and human languages. The sentiment of subjective elements in the data when we are doing sentiment mining as a part of natural language processing has to be identified. Sentiment analysis is also called as opinion mining, opinion extraction, sentiment mining [18]. This data for sentiment mining can be accumulated from various places on the internet which hosts input directly from the users. This can be comments about products on e-commerce sites, blogs, articles or even data from social networks. It is estimated that 84% of information technology processes deals with unstructured data [19]. This data can be used to improve the quality or products and services to satisfy the consumers in a better way. The sentiment analysis deals with determining type of emotion the data wants to convey. The opinion that is collected from the vast amount of data can range from extreme hostility to extreme appreciation or love. Positive, negative or neutral can be seen as the polarity of the content that is to be set for opinion extraction. Polarity can be obtained after completing all

the processes of opinion extraction. This data can be accumulated from social media websites where petabytes worth of data is consumed. From a financial perspective the opinion mining and sentiment analysis holds extreme value because it is estimated that 81% of people have gone through the consumer review of the goods they wanted to buy [17]. It is also estimated that a consumer is ready to give from 20% to 99% more for an item that has better ratings [13].

Sentiment analysis is that branch of the natural language processing which is used to identify and give out the polarity knowledge. Sentiment analysis can be used on different kinds of data where we want to find the mood of a human being or the views regarding a particular entity. A set of procedures which are used to know and to describe how a particular person thinks are included in sentiment mining. Opinion evaluation aids in deciding upon the views and opinions of the concerned person with respect to a few discipline. Perspective is also the choice of the concerned person to estimate, the opinionated data on the basis of the emotional state of the concerned person at the time when the data was recorded. Opinion extraction can also aid in investigating sentiment on different styles written records. It's going to rank the whole report as optimistic or negative, and it's going to additionally rank the isolated responses of person through phrases in the given dataset. Sentiment mining can aid in watching for a precise issue of interest, many organizations take the help of it to keep a look out for their products and services. Taking an illustration, if a large group of angry customers start mud-slinging on your brand on social media, then sentiment evaluation will be rate the post about the product as terrible and your product will be flooded with bad sentiment scores.

These days, Opinion extraction performs an essential role where more than a few computing devices are trying to find out the process that is being utilized for deciding on the sentiment of gigantic amounts of textual content or speech. The sentiment of people also is of great value during the elections as the sentiment can prove detrimental for any political party. In case of any restaurant the review given by users online across different categories like ‘nice of meals’, ‘services’, ‘dwelling room’ and ‘services’ supplied can improve the customer incoming.

Since data of enormous quantity is being shared online, it is easy to comprehend the need for sentiment evaluation as the large amount data cannot be processed manually.

The trouble is that there is not a single algorithm out there which can be 100% dependable. To separate heart-warming and hopeful viewpoint from gloomy and hopeless viewpoint can be challenging. For this many solutions have been put forth but only with efficiency of 80%

accuracy. The factors we should be always aware of regardless of the tool we are working on while analysing are:

a) **Manner of the context**: A word can have good indication or bad indication but it does not entirely depend on the word but the context in which the word is used. For e.g. If there is person who wants to commit suicide and posts it online then the comment" Great idea" is not a positive sentiment.
b) **Obscurity**: Human language is a very complex thing so many times a good or a bad word in a sentence does not represent any polarity of the sentence.
c) **Sarcasm**: Sometimes a witty sentence can have a meaning which is entirely different from the type of words which are being used in the sentence.
d) **Language**: As many global languages are evolving and new words are being added to the dictionary every year. Words can sometimes be used in a different context to what its dictionary meaning actually is.

1.2 Methods of Sentiment Analysis

Sentiment mining can be categorized into the following groups:

a) **Spot the keyword**: In this classifying method the unambiguous words such as joyful, worried, scared, and tired of something are classified as a part of text based categorization.
b) **Affinity**: It aids in detection of affecting observable words, but it also commissions subjective words with a likely "affinity" for specific sentiments.
c) **Statistical leverage**: It works upon on machine learning entities that includes the semantic latent analysis.
d) **Concept based leverage:** It works upon the important factors that consider data representation and therefore are used for detecting semantics which give a definite meaning. [4]

1.3 Different levels of Sentiment Analysis

Different levels for sentiment are as follows [21][22]:

A. Word level: The steps are as following:

a) Detecting and extrapolating attributes from user.
b) Finds out if the given attributes of a particular object are good, bad or neutral.
c) Assemblage of same attributes: yielding the essence of sentiment of an attribute based

on more than a single review.

B. Sentence level: The steps are as following:

a) Identify subjective or opinionated sentences
b) Different Classes may be objective and subjective.
c) Sentiment classification on each sentence.
d) Different classes may be: good, bad and neutral.
e) Assuming a review may contain just a single sentiment is not true all the time.

C. Document level: The steps are as following:

a) Sentiment mining based on reviews for a specific subject.
b) The classes can show different polarity being good, bad or neutral
c) Believing that the entire data is based upon a single opinion that can also be accumulated from the entirety of the data is not true.

1.4 Applications of Sentiment Mining

a) Buying goods online: When we buy anything from an e-commerce site whether it is a Mobile phone or cloths, the previous consumers who have already bought those things can post their review online. This review can be on different criteria's of the product that the user wants to buy. Sentiment mining can help the upcoming customer to evaluate the reviews of the customers so that in future all the good and bad qualities of the product is well understood by the analysis of those reviews.

b) Betterment in the goods and services: Using sentiment mining companies get directly get feedback from users and they can evaluate this feedback so that they can improve their service.

c) Research for future assignments**:** The results of the sentiment mining by a company can give new insights into the future plans of a company so that they can tap more funds inti the projects which were more liked by their customers.

d) Endorsement mechanisms: If the viewpoint of people on a subject is categorized as good, bad and neutral, we can comprehend which service or product would be liked by certain individuals just like the Google ads use the feedback of users to determine which ads to show to a specific number of users.

e) Detect "fire": Any mode in which users can write their viewpoints can become a source of huge conflict regarding to the opinions of people. Sentiment mining can detect use of profanity or angry words in different kinds of sources [20].

f) Detection of spam: There is a lot of misleading information on the internet and the many times this information is forced upon the user. Sentiment mining can find which information is spam and help the user to automate its removal.

g) Constructing policies: With the help of sentiment mining the politicians can take the opinion of citizens into consideration.

h) Improving elections: The election process largely depends on the sentiments of people. If the data is used to evaluate these sentiments and analyze the public perception strengthening democracy and creating better decision support systems.

1.5 Problems in Sentiment Mining

Since we are being bombarded with data all the time the data can be collected from various sources but the biggest difficulty in acquiring data from different sources is the language barrier. If data is written in a common language like English the extraction of valuable information is much simpler because of the wide availability of the published material for English. Although even if it's written in English there are a lot of challenges to be faced. Issues like a sarcastic sentence where the meaning can be exactly different than what a user wants e.g. If a user writes "this product is not so bad" and the algorithm has been designed to pick out positive or negative words like "good" and "bad" as a main criteria for Opinion extraction, the result will entirely wrong. The use of internet slang like "lol","2g","dope","g8","bs" can have a great effect of the meaning of a sentence and since new slangs become trending very quickly, it is very hard design a model which will keep in pace with these changes. Many times users like to show their feelings by expanding the syllables in a word e.g. if a user is over excited the user might write "grrrreeeeaaaatttttt" instead of simply "great". The important thing in sentiment mining is that many times users don't write about the desired topic but they comment about something else entirely e.g. If there is a video on Youtube showing product review and on the comment section people don't actually talk about the product but some other thing like the audio or video quality of that video or even criticizing or praising the make or the video. A lot of data online is not in English since other major languages in the world like mandarin, Spanish and Arabic are being widely used in the various parts of the world .Other than major languages spoken in the world minor languages of in different parts of the world hold a big chuck of data comprising of native languages or even mashup languages. The wide variety of languages hold a big problem for Opinion Mining and Sentiment Analysis since they cannot be directly translated

into a major language like English because many words might get distorted in the translation and hence changing the intent of the sentence entirely. The coming of android devices in the late 2000's has commenced a boom in the mobile industry and with the wide use of emoticons in messaging shows a better way of expressing feelings of the user about a particular topic. This leads to another dimension in Opinion Mining and Sentiment Analysis as now not only words or sentences but also emoticons can have a great impact on the meaning of a sentence. Since every emoticons represents a particular emotion it can be directly considered a specific word in the sentence. The words and emoticons in a particular sentence together give it a meaning and must be used in coherence.

1.6 Data Mining

Data mining process is a step by step procedure to find useful knowledge in the bulk of information. There are many data mining techniques which are used like clustering, classification, evaluation, language processing and so on. Classification process is one of the most common process. As classification is a supervised process it means that it aids in putting a class label on a set of non-classified tuples. Data miming can help a lot of corporations to make better products and services and to increase their revenue by looking for the patters of their profits. Many multinational companies have used data mining in the field of food items, clothing brands, watches, and electronic digital gadgets

1.7 Naïve Bayes Classifier

Among the various classification algorithms Naive bayes is one of the most commonly used classification algorithm. Naive bayes algorithm is based on the bayes theorem on conditional probability. In sentiment analysis it is widely used because it pretty simple to apply and the results are fairly good.

1.8 Maximum Entropy (ME) method

This algorithm is used in sentiment mining because it gives good results. The maximum entropy method is used when the training set is used to set different constraints on distribution. To use maximum entropy we need a set of attributes to be selected from the data so that analysis can take place.

1.9 Gradient Descent Method

This method is used when we are dealing with a pretty large dataset. In this method weights are assigned and after each round or iteration the weights in the data that we are training are modified.

CHAPTER 2

REVIEW OF LITERATURE

LITERATURE SURVEY

Asha S Manek et al. performed a look at on "Box-office Forecasting primarily based on Sentiments of Movie Reviews and Independent Subspace Method" .This paper proposes a way the usage of weight by using Gini Index technique for function selection and use of guide vector machines for sentiment analysis is used for prediction the usage of numerous massive movie statistics set is used. This consequences of the use of the gini index primarily based approach indicates better performance in phrases of accuracy and blunders charge [2].

Cagatay Catal et al. conducted a study having used multiple classifiers. In this paper vote algorithm is utilized by combining the 3 classifiers naïve bayes, help vector machines and bagging. The use of ensemble leaners approach is used right here. The use of combination of algorithms is better than using a single algorithm [8].

Abinash Tripathy et al. carried out a study using N-gram technique. This paper proposes to use different classification algorithms. N-gram means using more than one word for analysis. This gives it an edge over those techniques where a single word is used [9].

Tobias Gunther et al. carried out an extensive study Gradient Descent method by using attributes of a language. This paper proposes an technique to predict the feelings of Tweets and SMS primarily based on supervised gadget gaining knowledge of strategies. [10].

Chee Kian Leong et al. carried out an extenstive study opinion extraction in SMS texts for coaching assessment. This paper proposes exploration of the ability application of sentiment evaluation for texts messages in teaching evaluation. The three ways which we are moved towards are: the base version, the ‘‘corrected’’ version where mistakes in the spelling are fixed and the ‘‘sentiment’’ version where the previous I fixed my making the opinion extraction to be formed [11].

Shunxiang Zhang et al. carried out a look at on "Sentiment analysis of Chinese micro-blog text primarily based on prolonged sentiment dictionary". This paper proposes a sentiment mining technique for Chinese micro-weblog textual content based totally on the sentiment dictionary. First, the sentiment dictionary can be extended by way of extraction and creation of diploma adverb dictionary, community phrase dictionary, negative phrase dictionary and

different relative dictionaries. Second, the sentiment price of a micro-blog statistics can be obtained through the calculation of the weight. Finally, the micro-weblog facts on a subject can be classified as high quality, bad and neutral sentiments [12].

Kamal Nigam et al. carried out an extensive study on the use of Maximum Entropy for classification of text. This paper proposes using maximum entropy techniques for opinion extraction in text. Maximum Entropy is used in the categorization of text by using the distribution technique. In experimental results on numerous text datasets the accuracy was somewhat close to Naive Bayes and it seems that Maximum Entropy is on occasionally notably better, but additionally occasionally worse [7].

Bo Pang et al. performed a detailed examination on Sentiment Classification and different types of Machine Learning methods". In this paper it proposes using three device mastering strategies. Different classification algorithms are used here and are in comparison through the use of components of speech tagging and using n-grams [13].

M.Geetha et al. carried out an extensive study on the link between consumers opinion and and the ranking received online. In this paper there is the establishment of link of the customer sentiments and the accommodations of the facility. This observation takes into consideration the customer satisfaction and reduces it to the polarity of emotions. The evaluation within the study identifies the link between the customer opinion about the facility and the ranking of the accommodation. The customer opinion about the facility explains huge gap in customer rankings across each of the in categories. The take a look at unearths that, whilst as compared with premium resorts, managers of finances lodges have to make their team of workers performance and lodge services higher [14].
Chetashri Bhadanea et al. conducted a study on "Sentiment evaluation: Measuring opinions". This paper makes a speciality of the distinctive methods which can be used for classifying a given piece of herbal language textual content in line with the feelings expressed in it. A set of strategies were carried out for thing class and polarity identity of product review the usage of machine gaining knowledge of (SVM) blended with area unique lexicons. The experimental effects indicated about seventy eight% accuracy. The paper believes that the use of a larger dataset of patron critiques available at the Internet will improve the scope and usability of this software [15].

Tirath Prasad Sahu et al. This paper analyzed the sentiment analysis on movie assessment through preprocessing of information, then completed characteristic choice method carried

out and comparison of different type strategies achieved. Highest accuracy turned into given by means of random woodland with an accuracy of 88.95% [1].

Prerna et al. have implemented a device for sentiment evaluation by using Supervised Learning. The classifier is developed and then the data is trained using the support vector machine. In this method when support vector machine have been used and then some data is labeled as a neutral data and then that data is kept for a label [3].

Bogdon Batrinca et al. states that in todays world there are many kinds of social media platforms. It suggest different challenges that are faced during sentiment mining. This paper carries out a survey on the different techniques that are used for sentiment mining in case of social media and also the different tools that are available for it [4].

Ana Mihanovic et al. examine sentiment analysison various devices in two exclusive bureaucracy. It uses Knime tool for analysis. It uses sentiment mining on the twitter data with the help of hadoop and Hbase. As analysis works on certain facts these facts are loaded in the Knime tool. There is a grade scope used so that phrases that are used can be graded. The evaluation of the grades is on the basis of the sentiment it indicates. To construct a dictionary for social media data is also complex because it includes social slangs and very crude sarcastic sentences.[5].

HarunaIsah et al. represents a way to gather reviews of customers which buy different medicinal drugs and beauty products. Proposed study is to measure what the general people think about the different brands and the products such as medicinal drugs and beauty products from those bands. The data is brought from social media websites like twitter and facebook to analyse sentiments. Classifications algorithms are used here along with lexicon analysis and machine learning approach. [6].

CHAPTER 3

PROBLEM DEFINITION

PROBLEM DEFINITION

Various forms of knowledge are produced from specific Social media companies that should be equipped and to observe person's perspective in the direction of products, objects, movie assessment etc. There are millions of users on social media giants like Twitter and Facebook. Apart from social media e-commerce sites also have millions of users. The sentiment analysis using different kind of reviews can give new insights into the business model that the different companies follow and make the company more profitable. The major issue with sentiment analysis is that the mood of the user cannot be known and it causes a big difference in the analysis of what the user wrote and what the user really meant. The problem with information is the attributes with a vast number of qualities. It is one-sided towards picking properties with an expansive number of qualities. This may bring about over-fitting (determination of a quality that is non- ideal for forecasting). The major downside of SVMs is that they can be painfully inefficient to train. It will not work effectively without kernel, using kernel makes it computationally expensive, hence performs slow. The proposed Gini index feature selection addresses the issues of uneven dispersion of earlier class likelihood and worldwide decency of an element in two phases. Initially, it changes the examples space into a component particular standardized specimen's space without bargaining the intra-class highlight appropriation. In the second phase of the structure, it distinguishes the highlights that segregate the classes most by applying gini coefficient of disparity.

CHAPTER 4

SCOPE OF THE STUDY

SCOPE OF THE STUDY

The scope of the study is to analyze various sentiment analysis techniques in data mining and to collect movies review from IMDB movies reviews repository for further processing of sentiment analysis. Many different kinds datasets are available for sentiment forecasting but neither the dataset should be too small that the analysis is not performed well nor the dataset should be so large the training phase and the classification phase takes a lot of time. Gini index based feature selection methodology is proposed in this study for sentiment analysis. After applying filtering technique to the raw data, feature selection is done using Gini index technique and results of this are given to the classification algorithm. Results of the proposed algorithm are compared based on Accuracy, Precision, Recall and Fmeasure.

CHAPTER 5

OBJECTIVES OF THE STUDY

OBJECTIVES OF THE STUDY

a) To study and analyse various sentiment analysis techniques in Data mining
b) To collect movies review from IMDb movies reviews repository
c) To propose a technique using Gini Index based feature selection and ensemble learning for sentiment analysis prediction.
d) To analyse and compare the results of proposed approach with the existing based on Accuracy, Precision, Recall, Fmeasure.

CHAPTER 6

PROPOSED RESEARCH METHODOLOGY

RESEARCH METHODOLOGY

6.1 Method used

A. Collection of crude information and afterward apply filtering methods to make that crude information into organized organization. For doing the grouping, Text pre-handling and highlight extraction is a preparatory stage. Pre-preparing includes 3 stages:

a) **Word parsing and tokenization:** In this stage, every client audit parts into expressions of any regular preparing dialect.

b) **Stop words disposal:** There are words in the dataset that contain very little valuable information. As by expelling them, the execution increments and analysis can become better.

c) **Stem:** It is a procedure where a word is reduced to its root word that can be decided by applying the stemming rules on that particular word. The main cause of doing this is to reduce the size of the massive dataset that is being used for analysis.

B. Gini Index based Feature Selection

It is used to as a splitting method. In this algorithm we gather data sets for testing. Let A be the data sample, m be the divided no of subsets, P be the probability and Ci be the different classes then

$$\text{GiniIndex}(A) = 1 - \sum_{i=1}^{m} P_i^{\,2}$$

At the point when the base of GiniIndex (A) is zero then it means that all the records have a place with a similar classification at this gathering; it demonstrates that the greatest helpful data can be acquired. Then at the point when every one of the examples of accumulation have a standard circulation to a specific classification, GiniIndex(A) achieves greatest, demonstrating the base valuable data got.for gini index the little contaminating influence is the higher is the quality. On the other hand,

$$\text{GiniIndex}(A) = \sum_{i=1}^{m} P_i^2$$

measuring the contaminating influence of characteristic classify method, the greater is the contaminating influence the higher is the quality of attribute.

C. **Classification:** The classification process is the most important process in sentiment analysis. The classification algorithms if used in cohesion work better than if they are used as single because every algorithm has its disadvantages and when different algorithms are used together there disadvantages is reduced. To further improve the design bagging technique can be used for a chosen algorithm. By using different algorithms in an ensemble learner method the performance increases but at the same time a lot of time is utilized for training. But since the main purpose is to build a batter performing design as compared to single performing algorithms The voting phase is used to choose the one with the best performance

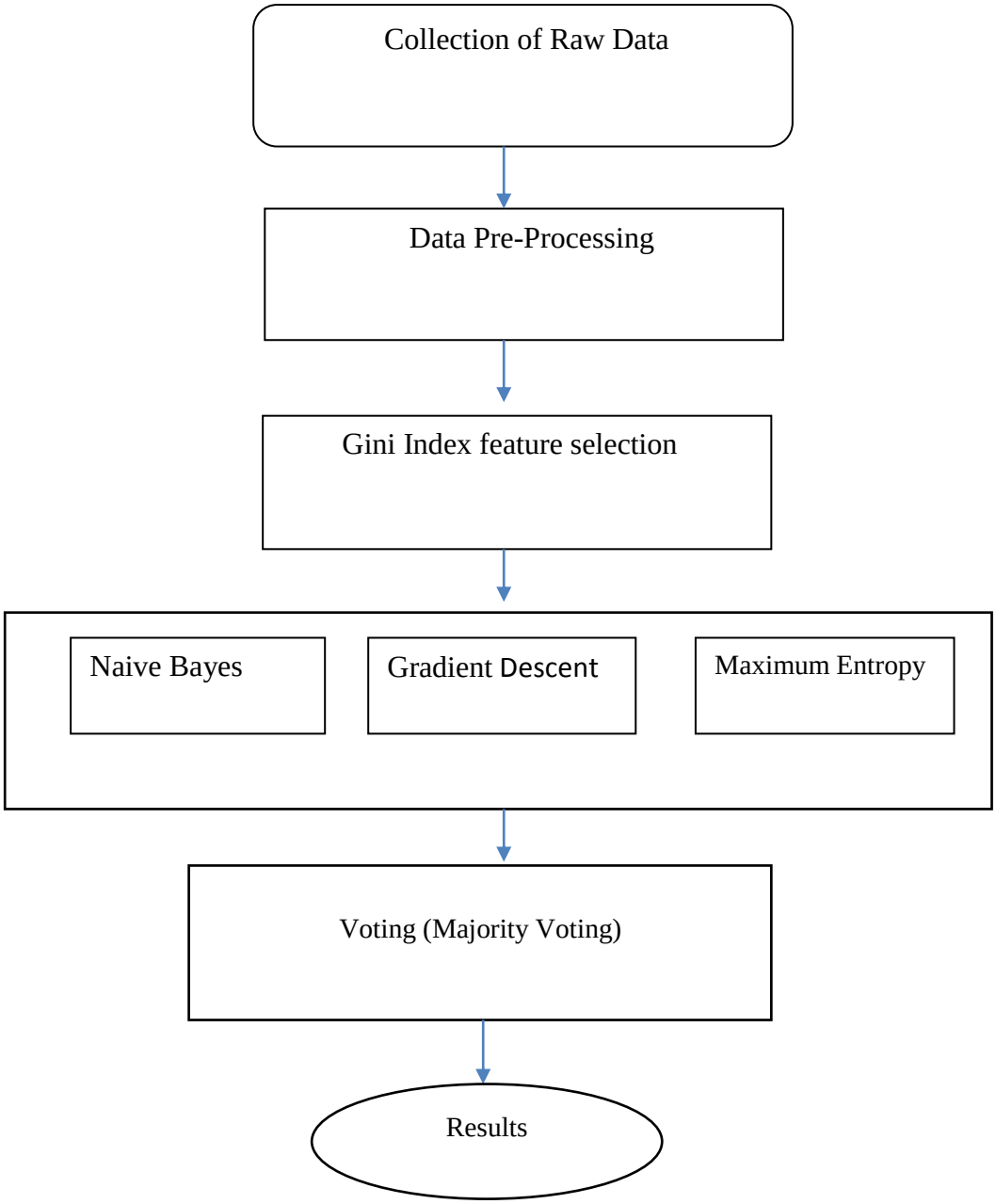

Figure 1: Flow chart of proposed methodology

CHAPTER 7

EXPECTED OUTCOMES

EXPECTED OUTCOME

With the outbreak of social media sites which are blooming with millions of users the need to evaluate this data is at its peak. The sentiment analysis can be used to evaluate the opinions of people. By design the sentiments are not well structured for directly being used for any form analysis but also a large part of information cannot be directly understood with using proper techniques. In general the sentiment of a user towards to subject of sentiment analysis can be categorized as compared to the polarity of sentiments the subject indicates.

Expected outcomes of this study is to propose a Gini index based feature selection and sentiment analysis method and to analyze and compare the results of proposed approach with the existing based on Accuracy, Precision, Recall, Fmeasure.

CHAPTER 8

SUMMARY AND CONCLUSION

Summary and Conclusion

The sentiment mining is becoming a booming part of the information technology industry and has attracted a lot of attention due to its applications. Since the previous feature selection techniques such as Term Frequency (TF) and Term Frequency-Inverse Document Frequency(TF-IDF) has its own disadvantages the use of Gini Index method for selection of features in analysis and using ensemble learner's technique to provide better results. The techniques that are used in the sentiment analysis might show different kind of results when using on different kind of datasets. Sentiment analysis as a part of natural language processing has a lot of applications which can be possible only when the problems that are faced while analyzing the data are solved. There can be use different algorithms which when used in a sophisticated manner will lead to better accuracy. There are different kinds of problems that need to be solved like identifying sarcasm, what a noun and pronoun is, multilingual approach, portmanteau analysis, what a phrase refers to and language issues.

CHAPTER 9

REFERENCES

REFERENCES

[1] Tirath Prasad Sahu and Sanjeev Ahuja, "Sentiment Analysis of Mvioe Reviews: A study on Feature Selection & Classification Algorithms", IEEE Xplore , 28 July, 2016.

[2] Asha S Manek, P Deepa Shenoy,M Chandra Mohan and Venugopal K R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier", Springer, pg 135–154, Feb 04,2016.

[3] Prerna Chikersal, Soujanya Poria, Erik Cambria (2015). "SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), 647–651.

[4] BogdonBatrinca, Philip C. Treleaven (2014) "Social media analytics: a survey of techniques, tools and platform",Department of Computer Science, Gower Street, London,UK published in Springer.

[5] Ana Mihanovic, HrvojeGabelica, ZivkoKrstic (2014) "Big Data and Sentiment Analysis using Knime: Online Reviews Vs. Social Media", MIPRO Opatija, Croatia

[6] HarunaIsah, Paul Trundle, Daniel Neagu. (2014)"Social media Analysis for product Safety using Text mining and Sentiment Analysis", Artificial Intelligence Research Group, University of Bradford, UK, IEEE.

[7] Kamal Nigam, John Lafferty, Andrew McCallum, "Using Maximum Entropy for Text Classification".

[8] Cagatay CATAL, Mehmet NANGIR, "A Sentiment Classification Model Based on Multiple Classifiers".

[9] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of Sentiment Reviews using N-gram Machine Learning Approach".

[10] Tobias Gunther ,Lenz Furrer, "GU-MLT-LT: Sentiment Analysis of Short Messages using Linguistic Features and Stochastic Gradient Descent".

[11] Chee Kian Leong, Yew Haur Lee , Wai Keong Mak, "Mining sentiments in SMS texts for teaching evaluation".

[12] Shunxiang Zhang, Zhongliang Wei, Yin Wang, Tao Liao, "Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary".

[13] Bo Pang and Lillian Lee, Shivakumar, “Thumbs up? Sentiment Classification using Machine Learning Techniques”.

[14] M. Geetha , Pratap Singha, Sumedha Sinha, “Relationship between customer sentiment and online customer ratings for hotels - An empirical analy”.

[15] Chetashri Bhadanea,Hardi Dalalb, Heenal Doshic, “Sentiment analysis: Measuring opinions”

[16] Waite Jeremy, Marketing Evangelist, Watson Marketing EMEA, 10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations,https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN

[17] Andreas Holzinger, Christof Stocker, Bernhard Ofner, Gottfried Prohaska, Alberto Brabenetz, and Rainer Hofmann-Wellenhof “Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field”

[18] Liu.B, “Sentiment Analysis and Opinion Mining”

[19] K. Douglas, “Infographic: big data brings marketingbig numbers,” 2012.

[20] Haseena Rahmath, “Opinion Mining and Sentiment Analysis - Challenges and Applications”.

[21] Mrs. R.Nithya, Dr. D.Maheshwari. (2014) “Sentiment Analysis on Unstructured Review”, International Conference on Intelligent Computing Application, IEEE, pp. 367-371, March 2014.

[22] Lukasz Augustyaniak, Tomasz Kajdanowicz, PrzemyslawKazienko, MarcinKulisiewicz, WlodzimierzTuliglowicz, “An Approach to Sentiment Analysis of Movie Reviews: Lexicon Based vs. Classification”, Springer, Vol. 8480, pp. 168-178, 2014.

CHAPTER 10

APPENDIX

Sentiment analysis	Technique used for analysis of a dataset for opinion forecasting.
Naive bayes	Algorithm used for classification.
Data mining	It is a process of finding useful information.
Polarity	It describes if polarity is positive or negative.
Classifier	Technique used for classification.
Opinion mining	The process of finding the opinion of a subject towards a target entity.