A Hybrid Approach for Enhancing Accuracy and Detecting Sarcasm in Sentiment Analysis

A Dissertation Proposal submitted

By

Shaina Gupta

To

Department of Computer Science and Engineering

In partial fulfillment of the Requirement for the
Award of the Degree of

Master of Technology in Computer Science and Engineering
Under the guidance of
Mr. Ravinder Singh
(November 2017)

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

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 :** CSERGD0338

Supervisor Name : Ravinder Singh **UID :** 17750 **Designation :** Assistant Professor

Qualification : M tech **Research Experience :** 4 years

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Shaina Gupta	11609008	2016	K1637	8699960899

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** Ravinder 17750

PROPOSED TOPIC : Opinion mining of news healdlines

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	7.50
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.50
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.	7.25
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.75
5	Social Applicability: Project work intends to solve a practical problem.	7.75
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.75

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): NA
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: **Opinion mining of news healdlines**

Overall Remarks: Approved

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

11/29/2017 10:02:13 PM

ABSTRACT

Sentiment analysis and opinion mining are gaining significant importance in the field of research. In last few years, many efforts had been made to mine the opinions and sentiments of people which they post on internet in the form of reviews, blogs, forum, discussions and micro-blogs. The important thing is to discover what people actually think. Their opinions are classified on the basis of polarity (positive, negative or neutral). In this study it has been discussed about what actually sentiment analysis is different technologies and different tools available for it and the areas in which work has been done has been discussed. Along with it proposed methodology has been discussed which is a hybrid approach which includes emoticon classification, Negation and Modifier classification, Hashtag classification, Sarcasm Detection, Polarity detection using SWNC and DSC.

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation/dissertation proposal entitled **A Hybrid Approach for enhancing accuracy and detecting sarcasm in Sentiment Analysis** submitted by Shaina Gupta at Lovely Professional University, Phagwara, India is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Ravinder Singh

Date:

1) Counter Signed by:

   Concerned HOD:

   HoD's Signature: _____

   HoD Name: _____

   Date: _____

2) Neutral Examiners:

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

3) Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

Gratitude cannot be seen or expressed. Often words are inadequate to serve as a model of expression specially the sense of indebtedness and gratitude to all those who help us in our duty. It is of immense pleasure and profound privilege to express my gratitude along with sincere thanks to my dissertation guide **Mr. Ravinder Singh** for granting me the opportunity to do research on the topic **A Hybrid Approach for enhancing accuracy and detecting sarcasm in Sentiment Analysis.** I am very much thankful to him for his co-operation and his assistance throughout my research work. He has been guiding light and without his guidance this report would not have been possible.

I am also gratefully indebted to **LOVELY PROFESSIONAL UNIVERSITY** for giving us such research opportunities time to time so that we can gain more knowledge. By participating in an effort like this I become aware of the degree to which other people supported me in by endeavor support me in cascades from family and friends. This completed research report is as much their achievements as it is mine.

Place: Lovely Professional University Shaina Gupta (11609008)

Date: 30-November-2017

DECLARATION

I hereby declare that the dissertation proposal entitled, **A Hybrid approach for enhancing accuracy in Sentiment Analysis** submitted for the 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.

Shaina Gupta

Regd. No: 11609008

Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 INTRODUCTION

From last few years, people are more dependent on internet for expressing their views. The expanding social networking, web has increased and people have started sharing data through various means. They share their data in the form of reviews, forums, discussions, blogs, micro-blogs. Emotions are associated with these data which can be positive, negative as well as neutral. Identifying these emotions through face to face communication is easy as compared to textual communication. But these days' social media have increased rapidly and a huge amount of textual data is available on internet which can be mined and managed so that people and organizations could benefit from it. We can gain understanding of the attitudes, emotions and opinions expressed with an online mention. Organizations and business are always concerned about what are people's opinions for their organization's product or services. So they can easily track their reputation in the media through text mining otherwise they have to undergo expensive surveys which may not provide true results always.

Even analysis of news headlines is also important as it tells weather news has a positive or negative impact on society. Rather than going through complete story, most people judge news contents directly by scanning news headlines. So even a small headlines plays an important role in any judgment. These headlines can be mined to detect the polarity as positive, negative or neutral.

1.2 SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is a text mining technique which is used to determine the opinions which are expressed by the author of the text. Opinion mining is useful in many

ways as in the marketing field it can help to detect the popularity of a particular product. Today even many newspapers are published online which are accessed through web or even news applications are also available where people post their opinions for that particular news. These opinions can be mined to detect positive, negative or neutral impact on the society. It is extremely useful to gain overview of public opinion behind certain topic. The components of opinion are:

i. **Opinion Holder:** It is defined as the holder of a particular opinion which can be an organization, a person that holds an opinion. In case of online blogs, Opinion holder is the person who writes these blogs or in simple words we can say that the one who gives the opinion.

ii. **Opinion Object:** It is defined as the object on which the opinion holder is expressing the opinion or in simple words we can say that on which opinion is given.

iii. **Opinion Orientation:** It is defined as the polarity or orientation of the opinion which can be positive, negative, or neutral.

Example: According to Alex this music system has excellent voice quality. In this example, Opinion holder is Alex, Opinion Object is voice quality and Opinion orientation is excellent which is positive.

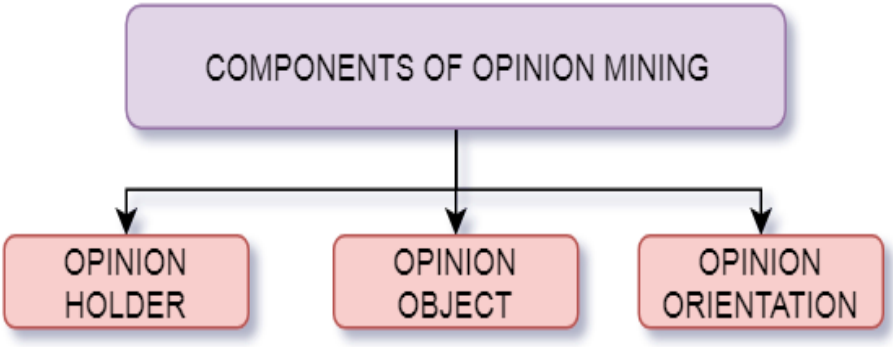

Figure 1: Components of Opinion Mining

1.2.1 LEVELS OF SENTIMENT ANALYSIS

Generally, there are three different levels of sentiment analysis which are as follow:

i. **Document Level**: Document level opinion mining is about classifying the whole document as positive, negative or neutral about a certain object. Assumption that is taken at this level is that each document focus on a single object and contains opinions from single opinion holder but it is not true in many cases.

ii. **Sentence Level:** In sentence level opinion mining, firstly the sentence is classified as subjective or objective and then the opinion of the sentence is analyzed. Assumption made in this is that one sentence contains only one opinion. It is not applicable for forum and blog as there could be multiple opinions on multiple objects.

iii. **Feature Level:** In feature level opinion mining, fine grained analyses are done to determine the opinion of an object i.e. positive or negative or neutral. It focuses on feature of a single object posted by a single opinion holder. It is also not applicable for forum and blog.

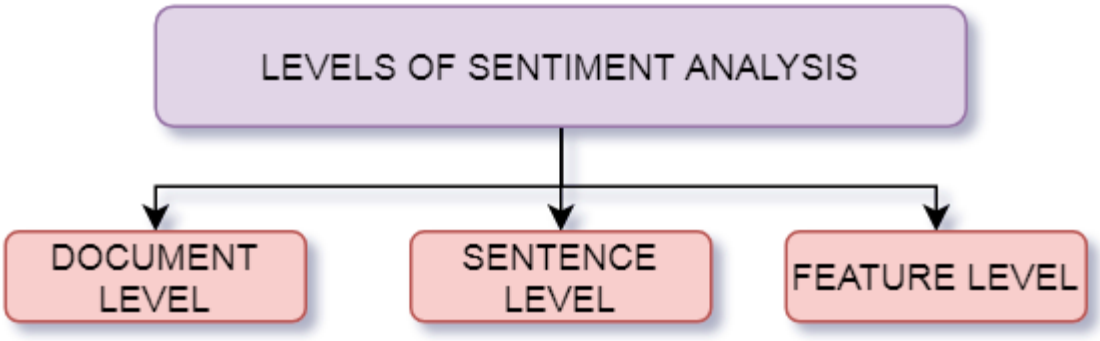

Figure 2: Levels of Sentiment Analysis

1.2.2 APPROACHES FOR SENTIMENT ANALYSIS

Basically there are two approaches for performing sentiment analysis which are as follow:

i. **LEXICON-BASED APPROACH:** Lexicon –based approach or rule based approach is most widely used in the field of sentiment analysis. Lexicon-based sentiment analysis associate with the presence of certain word in the document. Lexicon contains different features including part of speech tagging of words, their sentiment values and the subjectivity of words etc. Advantage of this approach is that it covers wider term but on

the other hand its limitations are that finite numbers of words are present in the lexicon and assignment of fixed sentiment orientation and score to words is there. It may be manually constructed, corpus based or dictionary based.

a) **SentiWordNet:** SentiWordNet is a lexicon resource for opinion mining. It assigns to each synset of WordNet which assigns it three opinions which are as positive, negative and objective. For each word it assigns a value ranging between 0.0 to 1.0. It basically performs three categories of task as subjectivity-objectivity Polarity, positivity-negativity Polarity and strength of positive-negativity Polarity. It assigns preprocessing steps like tokenization, POS tagging, stemming or normalization before calculating polarity. For combining scores there are different methods available like sum of all scores, average of all scores, sum of scores for adjective, average of scores for adjective, average of non-zero scores and majority votes.

b) **Emoji Sentiment Ranking:** It is a lexicon of 751 different emoji characters with automatically assigned sentiment scores. This sentiment is computed from different 70,000 tweets which are labeled by 83 different human annotators and are present in 13 European languages. It is used for sentiment analysis and scores are based on number of occurrences and their position in the sentence.

ii. **MACHINE LEARNING APPROACH:** Automated or machine learning based sentiment analysis requires the creation of a model by training a classifier. For that firstly we need a training data set with positive, negative and neutral classes extract features from the dataset and then train the algorithm based on our examples. Advantage of this is the ability to adapt and create trained models for specific contexts and purposes. But on the other hand limitation is that it has low applicability to the new data because it is necessary the availability of labeled data could be costly.

a) **Naïve bayes Classifier:** It is a supervised classifier given by Thomas bayes. This algorithm is implemented to calculate the probability of a data to be positive or negative.[1] Advantage is that the model is easy to interpret and efficient computation is there. It works best with textual data as well as numeric data formats. But on the other hand assumptions of attributes are made which may not be valid always.

b) **Support Vector Machine (SVM):** It is a supervised learning model which is based on the decision planes that defines the decision boundary.[2] This decision plane separates groups of instances that are having different class memberships. There are further extensions of SVM as Soft Margin Classification, Non Linear Classification, and Multiclass SVM. Advantage of this method is that its performance is very good and there is low dependency on data set. But on the other hand in case of missing value, it needs to be pre-processed and there is difficulty in interpreting the resulting model. SVM helps many researchers to perform short text classification as compared to full text.

c) **Multilayer Perception (MLP):** It is a feed forward neural network with 2 or more layers as input, output or hidden layers. Advantage of this method is that it does not enforce any sort of constraints with respect to initial data and it is a robust method when it deals with the problems containing noisy data. But on the other hand it needs more time to execute as its flexibility lies in the need to have enough training data.

d) **Clustering:** It is an unsupervised learning method and is a process of organizing objects and instances in a class whose members are similar. There are number of clustering algorithms that are becoming popular as Exclusive Clustering, Overlapping Clustering, Hierarchical Clustering, Agglomerative Clustering, Divisive Clustering, and Probabilistic Clustering. Advantage of this is that number of groups is generally known and there is no learning set of labeled observations.

e) **Sentic Computing:** It is a technique which enables computers to understand human emotions. For this computer needs conceptual information and relative affective information associated to it. It creates a database known as SenticNet.

f) **Decision Tree:** It is represented in the form of a tree and is used for text categorization where each node acts as a leaf. Decision tree is easy to understand and rules can be easily generated through them. These can solve complex problems very easily but training through them is very expensive. In case of news headlines analysis, news headline can only be connected to one branch which makes it unsuitable in this case.

g) **Maximum Entropy:** Maximum Entropy (MaxEnt) classifier is very closely related to Naive Bayes classifier, except that rather than allowing each feature to have its say

independently, the model uses search-based optimization to find weights for the features that maximize the likelihood of the training data.

h) **HYBRID APPROACH:** This approach uses both machine learning and lexicon based approach. Its advantage is that lexicon symbiosis, the detection and measurement of sentiment at the concept level is done and there is lesser sensitivity to the changes in the topic domain but it has noisy reviews.

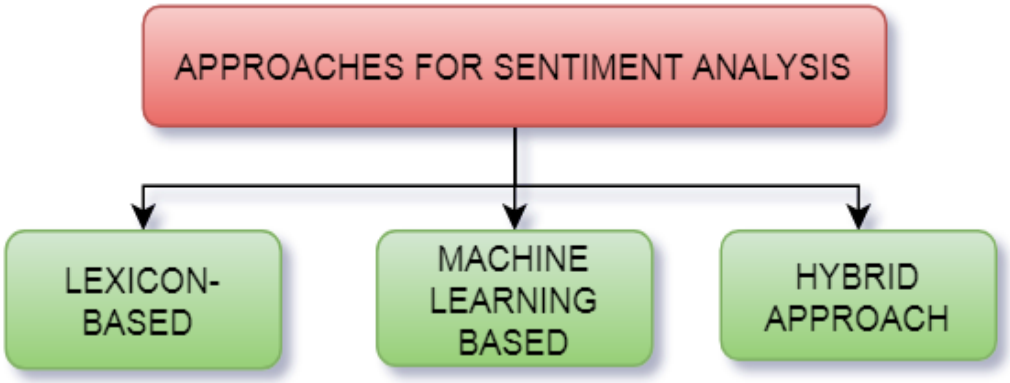

Figure 3: Approaches of Sentiment Analysis

1.3 CHALLENGES FACED IN SENTIMENT ANALYSIS

There are many challenges being faced in sentiment analysis which are described as follow:

i. **Sarcasm:** People not always express their opinions in the same way. Opinion of every individual people is different because of the way of thinking. Some people express it in sarcastic way that seems to be positive when read but are actually not. Remarks made by them mean the opposite of what they say. [3] They are usually made in order to hurt someone's feelings or to criticize something in a humorous way. Example: My flight is delayed…amazing…!

ii. **Spam and Fake reviews**: Usually web contains a lot of spam contents which should be avoided for proper analysis otherwise they can greatly affect the polarity of the text.[4]

iii. **Short forms:** Nowadays almost all prefer to use short forms like for how are you it is written as hru. It is the biggest challenge as it is difficult to interpret at machine level.

iv. **Negation Handling:** Negation words are those which reverse the polarity of sentence if they occur in sentence.[5] These words include not, never, can't, couldn't etc. So they must be properly handled in order to attain accurate polarity detection. Example: This smart phone is not good. In this example not totally reverses the polarity of the sentence from positive to negative.

v. **Domain dependency:** It is there as one word can be positive for one domain but negative for other domain. There are cases when one word has positive orientation in one domain and the same word has negative orientation in different domain[6]. Example: Maruti's small car has low fuel consumption and small size. In this example low describes about fuel consumption ad small is related to size but not to maruti. Similarly, if we say high fuel consumption and high chassis, in this case high has negative polarity in fuel consumption case but positive in case of chassis domain plays role in detecting the accurate polarity of the word.

vi. **Emoticons:** Emoticon symbols present in text also add to the value of polarity of sentences and they should not be ignored. Emoticon convey message more effectively without having dependency on the language and specific domain.[7] They have become a vital part of social media chat and public reviews these days. So, their detection and classification have become necessary for the development of efficient analysis application.

vii. **Modifier words:** There are some words which either enhance or reduce the polarity strength of sentiment words like very, pretty, even, few etc. Example: The medicine is so far very good. Here very is an enhancing modifier which is enhancing the polarity of good.[8]

viii. **Hashtags:** Hashtags are usually removed while preprocessing but words written with hashtags put more emphasis and affects the polarity of the word.

1.4 TOOLS USED IN OPINION MINING

There are different tools available for tracking the polarity of user's opinions:

i. **SentiWordNet 3.0:** It is used for tokenization, POS tagging, Normalization, polarity assignment.

ii. **Stanford Corenlp**: It is used for POS tagging, Named Entity Recognizer, Parsing, and Sentiment Analysis etc.

iii. **Weka:** It is a machine learning algorithm used for data mining, Data pre-processing, Classification, Regression, Clustering, Association Rules, and Visualization.

iv. **NTLK:** It is used for Classification, Tokenization, Stemming, Tagging, Parsing, Semantic reasoning and also provides lexical resources such as WordNet.

v. **Rapid Miner:** It is used for data preparation, deep learning, machine learning, predictive analysis and text mining.

vi. **Review Seer Tool**: It is used to automate the work that is done by aggregation sites. In this, Naïve Bayes classifier is used to collect positive and negative opinions and that is done by assigning a score to the extracted feature term.

Chapter 2

REVIEW OF LITERATURE

Farhan Hassan Khan *et al (*2014) have used hybrid approach for classifying data of twitter. This hybrid algorithm included Emoticon Classifier (ECC), Improved Polarity Classifier (IPC), and SentiWordNet Classifier (SWNC). It has precision of 85.3%, Recall of 82.2%, Accuracy of 85.7%. EEC gave 70% accuracy but it is not god to use when there is no emoticon present. IPC has classified text as positive, negative and neutral by calculating the sum of polarities by using SWNC. It has resolved the data sparsity issue using domain independent techniques. Preprocessing, EEC, IPC and SWNC play a major role in resolution of sparsity issue.[7]

Muhammad Zubair Asghar *et al* (2017) have performed a lexicon enhanced Sentimental Analysis in which they have dealt with issues like domain dependency, emoticon analysis, negation handling, low accuracy of the classifiers due to modifiers and negations. For emoticon Classification they have used enhanced emoticon classier which contained 230 different emoticons with their respective polarities. This classifier has an accuracy of 91.2% with kappa score of 0.85 which is quite high. For modifier and negation handling they have used manual methods. Sometimes polarity of words that are not available in SWNC can be calculated by using domain specific classifier (DSC). If both SWNC and DSC give same result, then any resultant polarity can be selected else result of DSC is considered to be more accurate. This proposed method has achieved precision of 0.83, Recall of 0.94 and F-measure of 0.85.[8]

Prashant Raina *et al* (2013) have used SenticNet for sentiment analysis of news articles. It presented an opinion mining engine which exploits common-sense knowledge that is extracted from concept Net and SenticNet. Sentic Computing technique is useful for fine grained sentiment analysis. Proposed engine has been tested on a large corpus of sentences from news articles. This method has precision of 91% for neutral opinions, F-measure of 59% ,66%and 79% for positive, negative and neutral sentences and accuracy of 71% in classification.[9]

Kai Yang *et al* (2017) have used hybrid model by combining SVM and GBDT for classifying text. SVM is suitable when sentences have simple structure and strong opinion tendency whereas

GBDT performs well for long sentences with many sentiment words. [6]It has addressed the problem of dependence i.e. sentiments are not always directly related to nearby entities but to the aspect of those entities. The proposed work deals firstly with finding out entities,then aspects of the entities, secondly sentimental words are used to describe the sentiments of aspects and finally sentiments of the entities is obtained by combining sentiments of their aspects together. This work concluded that hybrid model outperforms the baseline models.

Alex M.G. Almeida *et al* (2016) *have* used four different hybrid approaches. Hybrid 1 approach used emoticon attribution, SWN, CESA, CBPI. Hybrid 2 is same as hybrid 1 but is different in CESA where it emphasis on negative message. Hybrid 3 has no baseline filters but work directly on CESA and CBPI. Hybrid 4 approach is same as Hybrid 3 but differs in CESA where it emphasis on negative messages. It concluded that Emoticon Analysis approaches i.e. Hybrid 1 and Hybrid 2 provides better result and hybrid approaches achieve better precision in case of neutral sentences. Hybrid 1 and Hybrid 2 achieved 76.99% and 77.78% while CESA approach achieved 70.66%. Hybrid 2 is ranked 1 followed by Hybrid 1 and then CESA.[10]

Simon Fong,Yan Zhuang *et al* (2013) have used Mallet as a tool and have compared different techniques i.e. Naïve Bayes, Maximum Entropy , Decision Tree ,C4.5 Decision Tree, Winnow and Balanced Winnow. It is found that Naïve Bayes performs the best of all six techniques while winnow are the worst techniques. Maximum Entropy classifier and C4.5 Decision Tree are second-best.[11]

Mazhar Iqbal Rana *et al (*2015) *have* discussed about text classification process, classifiers like Naïve Bayes, Support Vector Machine (SVM), Artificial Neutral Network and Decision Tree and other numerous feature extraction methods. Along with that detail of preprocessing has been discussed in which different steps of preprocessing has been discussed like tokenization, stop word removal, stemming etc. Different stemmers are available like S-Stemmer, Lovins-Stemmer, Porter-Stemmer, Paice/Husk Stemmer out of which Porter-Stemmer.[12] Different feature selection methods are also there like Boolean weighing, class frequency threshholding, and term frequency inverse class frequency and information gain. It concluded that different classification scenarios and algorithm perform different depending on news and data generated.

Apporv Aggarwal *et al (*2016) have proposed an algorithm which classifies the given news headlines whether they have positive impact or negative impact on society. For evaluating any

headline, they have used three algorithms: Algorithm 1, Algorithm 2 and Algorithm 3. Algorithm 1 is used to preprocess the words that are taken from News Headline. While preprocessing it has made use of POS-*Tagger, Lemmatization, and Stemming* steps. Algorithm 2 is used to analyze a news headline. SentiWordNet 3.0 has been used to identify the positive and negative score of every word which evaluates the total positive and negative impact in that news headline**.** In this method they have analyzed each and every word in the news headline whether it is a noun, verb, adverb, adjective or any part-of-speech. Algorithm has been experimented on 500 news headlines of past 30 days and it has provided deviation for 5 days as 3.2 but when number of days are increased, the deviation decreased due to sarcasm present in headlines.[3]

Vibha Soni *et al* (2014) *have* done opinion mining and have worked on aspect level analysis using SentiWordNet. They have used two methods in SentiWordNet which are Adverb-Adjective combination and Adverb-Adjective & Adverb–Verb combination and concluded that both methods are equally good and provide similar results. Work has been done on mobile review from Amazon.com. For implementation they have used java Net beans IDE. [13]

Priyanka *et al (*2013) have worked on best feature extraction for sentiment classification in which word Polarity Group and Word Polarity Sum are two different feature extraction methods out of which Word Polarity provides better result when combined with POS. For feature Pruning Subjectivity Scores and Proportional Difference are different methods and later provides results with high accuracy. It has been tested on online customer review and in this SVM is used as classifying algorithm. Their model had produced a highest accuracy of 91.9% among the entire feature combination when it has been experimented on small dataset and a accuracy of 95% when experimented on large dataset.[2]

Tanvi Hardeniya *et al* (2016) have used SentiWordNet to classify text as positive or negative.[5]Major issues addressed in this paper is negation handling and domain dependency. Work has been done mainly on negation handling which is done by using Fuzzy Logic. Amazon data set has been used for verifying the result.

Parvesh Kumar *et al (*2014) *have* discussed about advantages and disadvantages of various opinion mining techniques like Naïve bayes, Clustering, Support Vector Machine(SVM), Multilayer Perception (MLP). It concluded that Naïve Bayes is best suitable for textual

classification while clustering for customer services and SVM for biological reading and interpretation.[14]

A'sin Seedahmed Ali *et al (*2015) have discussed about the overview of opinion mining and sentiment analysis in detail along with different techniques that are available for it. [15] They have also provided the architecture of opinion mining which included opinion retrieval, opinion classification and opinion summarization.

Thakare Ketan Lalji *et al (*2016) have used both lexicon based approach and machine learning approach and have discussed about lexicon based approach that has high precision and low recall and learning algorithm like SVM has been used as a combination to increase the performance for classifying twitter messages. Negation handling has been done by replacing word with! Accuracy has been checked against classification features like Unigram, Bigram and Trigram with different training sets and concluded that with the increase in data sets accuracy is also increased.[16]

Shweta Rana *et al (*2016) *have* used Rapid Miner as a tool for mining and have worked on the comparison of Naïve Bayes and Support Vector Machine (SVM). It concluded that movie drama has high accuracy rate among different genre of the movies. Linear SVM has provided best accuracy followed by the synthetic words approach. Data set on which work is done is internet movie reviews.[1]

Asmita Dhokrat *et al* (2015) has taken into account the basic requirements of opinion mining to explore the present techniques used to develop full-fledged system.[17] It also list down various tools that are available for opinion mining like WEKA, NLTK, STANFORD CORENLP and different challenges being faced in it like domain independence, detection of spam and fake reviews ,use of abbreviations and short forms.

Nidhi Mishra *et al* (2012) *have* focused on the classification of opinion mining techniques that conveys user's opinion that may be positive, negative or neutral at various levels. Spam filtering issue has been mentioned in this which is a research issue in opinion mining.[4]

AUTHOR	Farhan Hassan Khan	Muhammad Zubair Asghar	Prashant Raina
YEAR	2014	2017	2013
TITLE	TOM: Twitter opinion mining framework using hybrid classification scheme	Lexicon-enhanced sentiment analysis framework using rule-based classification scheme	Sentiment analysis in news articles using sentic computing
TECHNIQUE	Hybrid Approach (EEC,IPC,SWNC)	Lexicon Enhanced Sentiment Analysis(EC,DSC,SWNC,MNC)	SenticNet
DATA	Twitter	Reviews	News Articles
PRECISION	85.3%	83%	91% (for neutral opinions)
RECALL	82.2%	94%	-
F-MEASURE		85%	59%(positive) 66%(negative) 79%(neutral)
ACCURACY	85.7%	-	71%

Table 1: Summary of Literature Review

Chapter 3

SCOPE OF STUDY

The expanding social networking, web has increased and people have started sharing data through various means. They share their data in the form of reviews, forums, discussions, blogs, and micro-blogs. Emotions are associated with these data which can be positive, negative as well as neutral. Identifying these emotions through face to face communication is easy as compared to textual communication. Sentiment analysis or opinion mining is a text mining technique which is used to determine the opinions which are expressed by the author of the text. We can gain understanding of the attitudes, opinions and emotions expressed with an online mention. Organizations and business are always concerned about what are people's opinions for their organization's product or services. So they can easily track their reputation in the media through text mining otherwise they have to undergo expensive surveys which may not provide true results always. Even analysis of news headlines is also important as it tells weather news has a positive or negative impact on society. Most people judge news contents directly by scanning news headlines rather than going through complete story. So even a small headlines plays an important role in any judgment.

Chapter 4

OBJECTIVE OF THE STUDY

This study aims to explore and analyze various sentiment analysis techniques and develop a hybrid classification technique in order to improve the accuracy of the detection of polarity of the text. New improved technique primarily focuses on the detection of emoticons present along with the text as they also convey sentiment, detecting sarcasm and dealing with other issues like negation and modifier words. It will be developed in such a manner that it deals with all these issues and provide us with the improved polarity value and orientation.

The research is focused on following objectives:

1. To detect sarcasm present in sentences and modify the polarity according to sarcasm.
2. To improve the emoticon polarity detection by using emoji sentiment ranking lexicon this is based on number of occurrences and position of the word.
3. To deal with the negation words by reversing the polarity of such words and modifying the polarity due to modifiers.
4. To modify the polarity of the words which are specific to a particular domain.

Chapter 5

RESEARCH METHODOLOGY

After review of approaches and techniques in the area of Sentiment Analysis it has been observed that there is more improvement required to achieve higher accuracy in determining the polarity or sentiments of the words. This can be achieved if we deal with major issues faced in opinion mining.

5.1 A FRAMEWORK FOR RESEARCH METHODOLOGY

Following Figure 4: Proposed Frameworks shows the framework of the approach which can be used to enhance the accuracy of sentiment analysis.

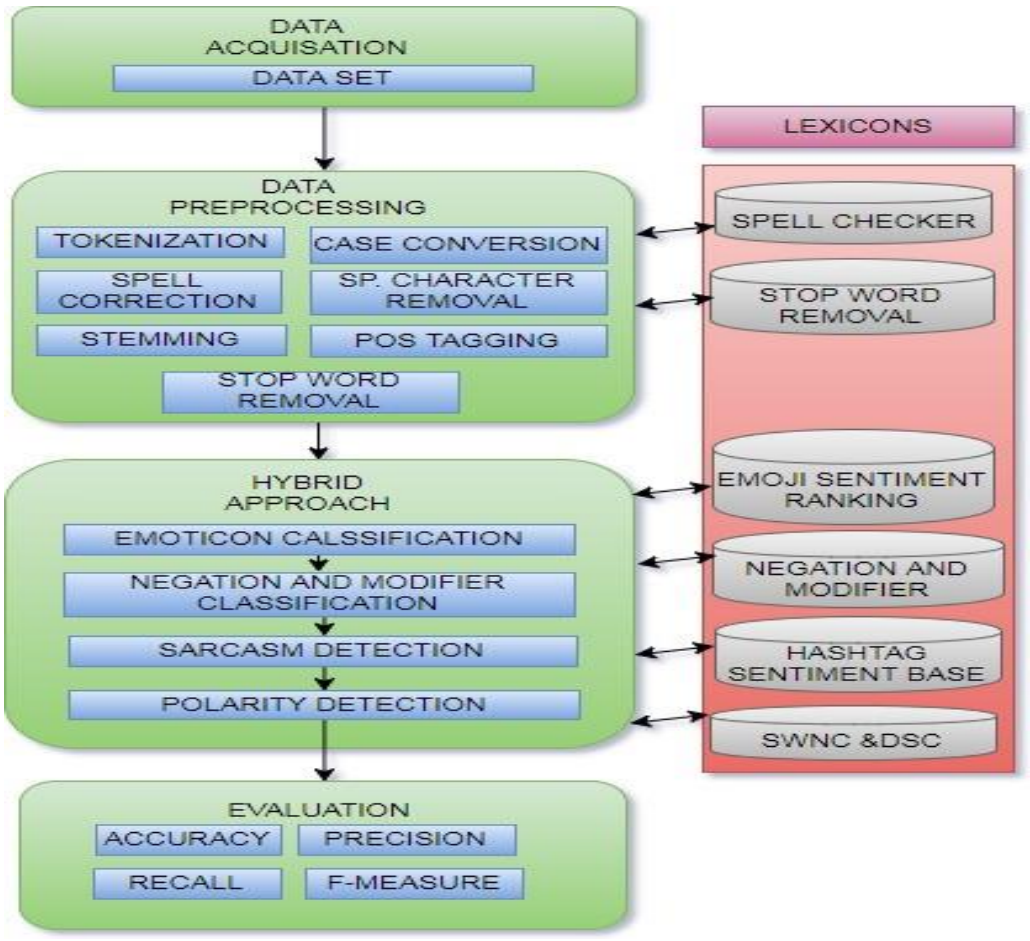

Figure 4: Proposed Framework

5.2 STEPS INVOLVED IN THE PROCESS

There are some steps to be followed as mentioned under:

5.1.1. Data Acquisition: Data can be collected from various sources which may include print media, World Wide Web, electronic media etc. Many researchers have classified news based on headlines by gathering headlines from RSS feeds of various websites using RSS bandit tool.

5.1.2. Data Preprocessing: Text preprocessing is major step in text mining. It further includes many processes like tokenization, stemming or lemmatization etc.

i. **Tokenization:** Tokenization is defined as breaking huge text into smaller tokens or segments which are served as input for further processing.

ii. **Case Conversion:** All the words are converted from upper case to lower case so that they can be easily processed further.

iii. **Spell Correction:** There are some words which are incorrectly spelled, so they are corrected in order to find the polarity otherwise they would be skipped and it might be possible that those words greatly affect the polarity.

iv. **Special Character Removal:** From tokens obtained after tokenization special characters like commas, quotes, semi-colons, full stops, underscores, dashes, brackets etc. are removed.

v. **Stop Word Removal:** Stop words are those which appear frequently in text and are considered of low worth and should be removed.

vi. **Stemming:** Stemming or lemmatization is the process in which word is reduced to its root word by removing affixes. There are many stemmers available like S-Stemmer, Lovins Stemmer, Porter Stemmer, and Paice/Husk Stemmer out of which Porter Stemmer is the most appropriate one.

vii. **POS Tagging:** Part of speech tagging is which assigns part of speech such as noun, verb, adjective, adverb to each word of the text. Further these tagged words are used for feature extraction and polarity calculation.

5.1.3. Hybrid Approach: In hybrid approach, there are different classifiers which are used to deal with different problems so that the polarity can be improved. These classification methods are:

i. **Negation and Modifier Classification:** Negation words are those which reverse the polarity of sentence if they occur in sentence. These words include not, never, can't, couldn't etc. So they must be properly handled for accurate polarity detection. Example: This smart phone is not good. In this example not totally reverses the polarity of the sentence from positive to negative. There are some words which either enhance or reduce the polarity strength of sentiment words like very, pretty, even, few etc. The words which increase the polarity strength of the sentiment word are called enhancer and those which reduce the polarity strength are called reducers. Example: The medicine is so far very good. Here very is an enhancing modifier which is enhancing the polarity of good.

ii. **Hashtag Classification:** Hashtags are usually removed while preprocessing but words written with hashtags put more emphasis and affects the polarity of the word. For this we have Hashtag Sentiment Base which contains the sentiment strength of the words with hashtags.

iii. **Emoticon Classification:** Emoticon convey message more effectively without having dependency on the language and specific domain. They have become vital part of social media chat and public reviews. So, their detection, classification has become necessary for the development of efficient analysis application. In earlier work, usually these emoticons were ignored. F.H. Khan has used emoticon classifier which contained 145 emoticons out of which 70 were categorized as positive and 75 as negative. Muhammad have used enhanced emoticon classifier in which it was extended from 145 to 230 set of emoticons out of which 120 were categorized as positive and 110 as negative. These both contain simple polarity but there is one more classifier named Emoji Sentiment Ranking which is a sentiment lexicon of 751 most frequent

used emoji and their polarity is based on the number of occurrences and position.

iv. **Sarcasm Detection:** People not always express their opinions in the same way. Opinion of every individual people is different because of the way of thinking. Some people express it in sarcastic way that seems to be positive when read but are actually not. Remarks made by them mean the opposite of what they say. They are usually made in order to hurt someone's feelings or to criticize something in a humorous way. Example: My flight is delayed…amazing…!. Recently a sentimental Analysis Technology has been developed which can detect sarcasm better than humans can by using emoji. Basically it is a model which is built by using deep learning. For a sentence it returns the emoji which is mostly used with that particular message.

v. **Polarity Detection:** Lexicon based approach is used to detect the sentiment of the text. Polarity Detection is done using SentiWordNet 3.0 in which each word. But there are some words which are not present in the SWNC and mostly words present in it are domain independent. So for dealing with domain dependency, we have Domain Specific Classifier (DSC). If both SWNC and DSC gives the same result, then we can pick anyone else the result of DSC is considered to be more accurate.

5.1.4. Evaluation: This is the final stage for Sentiment Analysis in which we determine how accurate is our proposed methodology. There are different performance metrics on basis of which evaluation is done which are as follow:

i. **Precision:** It is defined as the ratio of correctly predicted positive observation to the total predicted positive observations. High precision means low false positive rate. It can be calculated by formula:

$$Precision = \frac{TP}{TP + FP}$$

Where TP are true positive observations and FP are false positive observations.

ii. **Recall:** Recall is defined as the ratio of correctly predicted positive observations to all observations in actual class. It is calculated by formula:

$$Recall = \frac{TP}{TP + FN}$$

Where TP are true positive observations and FN are false negative observations.

iii. **F-Measure:** It is defined as a measure that combines precision and recall and is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2.\frac{Precision\ .\ Recall}{Precision + Recall}$$

iv. **Accuracy:** Accuracy is considered as the most intuitive performance measure and is a simple ratio of correctly predicted observation to the total observation. It is calculated as follow:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Where TP are true positive observations, FN are false negative observation, TN are true negative observations and FP are false positive observations.

5.3 FLOWCHART OF PROCESS

5.3.1 HASHTAG, NEGATION AND MODIFIER CLASSIFICATION

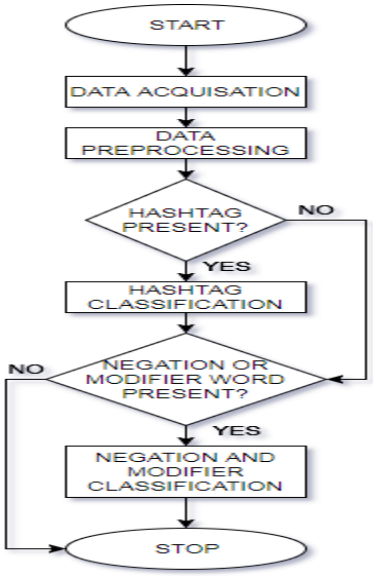

Figure 5: Flowchart for Hashtag, Negation and Modifier Classification

5.3.2 EMOTICON CLASSIFICATION AND SARCASM DETECTION

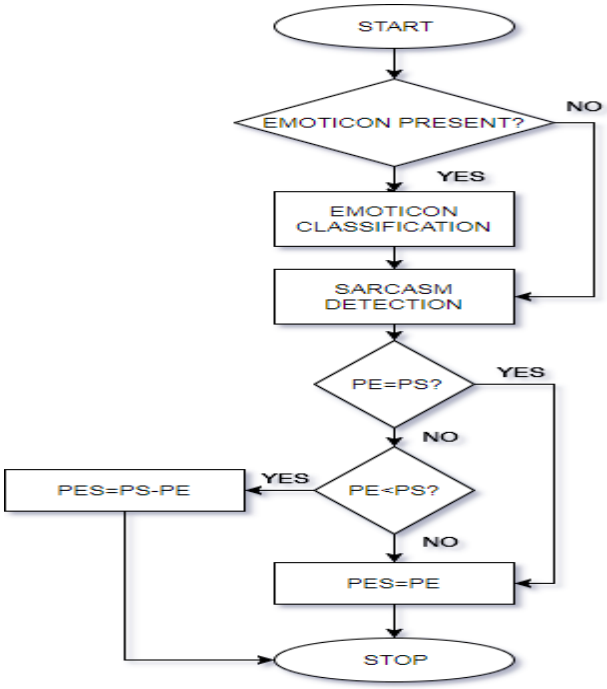

Figure 6: Emoticon Classification and Sarcasm Detection

Where **PE** = Polarity of emoji

PS = Polarity of sarcasm emoji

PES = Resultant polarity

5.3.3 POLARITY DETECTION

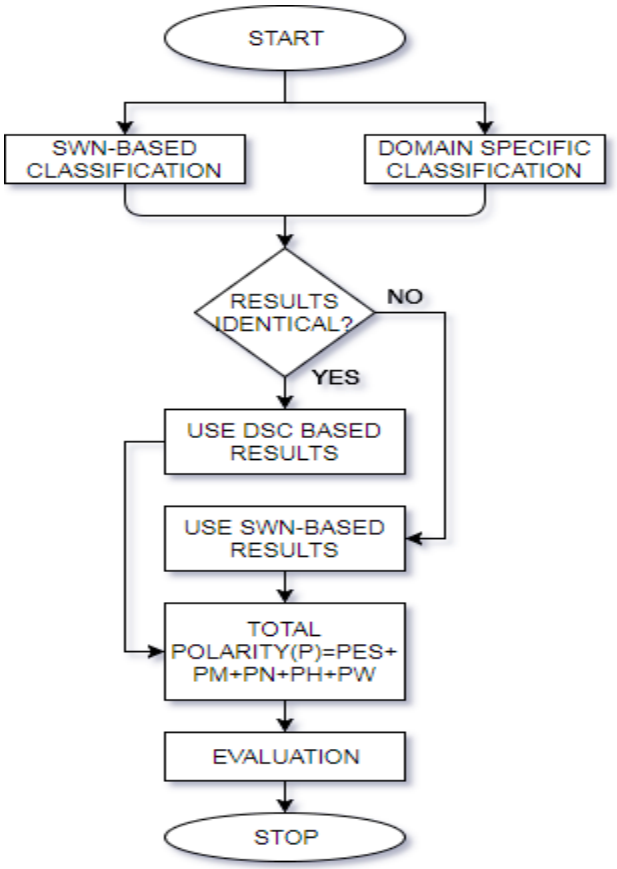

Figure 7: Polarity Detection

Where **PM** = Polarity of modifier

PH = Polarity of hashtag

PW = Polarity of word

PN = Polarity of negation

PES = Resultant polarity

Chapter 6

EXPECTED OUTCOME

The expected outcomes after the implementation of this methodology could be summarized as follow:

i. Detection of sarcasm that is generally present in reviews, comments or news headlines.
ii. Improved emoticon classification with the help of advanced emoji sentiment ranking algorithm.
iii. Improved accuracy of polarity detection by considering domain dependency, negation words, modifiers, and hashtags.

Chapter 7

CONCLUSION

These days' social media have increased rapidly and a huge amount of textual data is available on internet which can be mined and managed so that people and organizations could benefit from it. We can gain understanding of the attitudes, opinions and emotions expressed with an online mention. Opinion mining or Sentiment Analysis helps us to detect the polarity of those texts. Accuracy of algorithm used for sentiment analysis can be improved by adopting hybrid approach in which we can use two or more approaches. Emoticon analysis also helps to increase the accuracy of the algorithm as these days' text if often accompanied with emoticons. Detection of sarcasm is also important as it may directly reverse the polarity. Generally news mining is the latest field for mining where we can analyze weather particular news is making a positive, negative or neutral impact on people.

REFERENCES

[1] S. Rana and A. Singh, "Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques," *Int. Conf. Next Gener. Comput. Technol.*, no. October, pp. 106–111, 2016.

[2] C. Priyanka and D. Gupta, "Identifying the best feature combination for sentiment analysis of customer reviews," *Adv. Comput. Commun. Informatics (ICACCI), 2013 Int. Conf.*, pp. 102–108, 2013.

[3] A. Agarwal, V. Sharma, G. Sikka, and R. Dhir, "Opinion mining of news headlines using SentiWordNet," *2016 Symp. Colossal Data Anal. Networking, CDAN 2016*, 2016.

[4] N. Mishra and C. K. Jha, "Classification of Opinion Mining Techniques," *Int. J. Comput. Appl.*, vol. 56, no. 13, pp. 975–8887, 2012.

[5] T. Hardeniya and D. A. Borikar, "An Approach To Sentiment Analysis Using Lexicons With Comparative Analysis of Different Techniques," *IOSR J. Comput. Eng. Ver. I*, vol. 18, no. 3, pp. 2278–661, 2016.

[6] Y. Kai, Y. Cai, H. Dongping, J. Li, Z. Zhou, and X. Lei, "An effective hybrid model for opinion mining and sentiment analysis," *2017 IEEE Int. Conf. Big Data Smart Comput. BigComp 2017*, pp. 465–466, 2017.

[7] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," *Decis. Support Syst.*, vol. 57, no. 1, pp. 245–257, 2014.

[8] M. Z. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. A. Khan, "Lexicon-enhanced sentiment analysis framework using rule-based classification scheme," *PLoS One*, vol. 12, no. 2, p. e0171649, Feb. 2017.

[9] P. Raina, "Sentiment analysis in news articles using sentic computing," *Proc. - IEEE 13th Int. Conf. Data Min. Work. ICDMW 2013*, pp. 959–962, 2013.

[10] A. M. G. Almeida, R. A. Igawa, E. C. Paraiso, and S. N. Moriguchi, "Opinion Mining : A Comparison of Hybrid Approaches," no. c, pp. 1–7, 2016.

[11] S. Fong, Y. Zhuang, J. Li, and R. Khoury, "Sentiment Analysis of Online News Using MALLET," *2013 Int. Symp. Comput. Bus. Intell.*, pp. 301–304, 2013.

[12] M. I. Rana, S. Khalid, and M. U. Akbar, "News classification based on their headlines: A review," *17th IEEE Int. Multi Top. Conf. Collab. Sustain. Dev. Technol. IEEE INMIC 2014 - Proc.*, pp. 211–216, 2015.

[13] V. Soni and M. R. Patel, "Unsupervised Opinion Mining From Text Reviews Using SentiWordNet," *Int. J. Comput. Trends ...*, vol. 11, no. 5, pp. 234–238, 2014.

[14] P. K. Singh and M. Shahid Husain, "Methodological Study Of Opinion Mining And

Sentiment Analysis Techniques," *Int. J. Soft Comput.*, vol. 5, no. 1, pp. 11–21, 2014.

[15] S. Seedahmed Ali, "Opinion Mining Techniques," *IJISET -International J. Innov. Sci. Eng. Technol.*, vol. 2, no. 6, pp. 752–755, 2015.

[16] V. Nandi and S. Agrawal, "Sentiment Analysis using Hybrid Approach," *Int. Res. J. Eng. Technol.*, pp. 1621–1627, 2016.

[17] Dhocart A. et al., "Review on Techniques and Tools used for Opinion Mining," vol. 4, no. 6, pp. 419–424, 2015.