HOLISTIC MULTILINGUAL SENTIMENT ANALYSIS ON REVIEWS IN SOCIAL MEDIA

A thesis submitted in fulfillment for the requirements of the degree of

Doctor of Philosophy

by

SUKHNANDAN KAUR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT, DISTRICT SOLAN-173234, H.P.
INDIA

January, 2019

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
(Established by H.P. State Legislative Vide Act no. 14 of 2002)
Waknaghat, P.O. Waknaghat, Teh.- Kandaghat,
Distt. Solan – 173234 (H.P.) INDIA
Website: www.juit.ac.in
Phone No. (91) 01792-257999
Fax: (91) 01792 245362

CERTIFICATE

This is to certify that the thesis entitled, **“Holistic Multilingual Sentiment Analysis on Reviews in Social Media”** which is being submitted by **Sukhnandan Kaur (Enrolment No. 136215)** in fulfilment for the award of degree of **Doctor of Philosophy in Computer Science and Engineering by Jaypee University of Information Technology, Waknaghat, India** is the record of candidate’s own work carried out by her under my supervision. This work has not been submitted partially or wholly to any other Institute or University for the award of this or any other degree or diploma.

Dr. Rajni Mohana
Associate Professor
Department of Computer Science and Engineering
Jaypee University of Information Technology
Waknaghat,Solan– 173234
India

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
(Established by H.P. State Legislative Vide Act no. 14 of 2002)
Waknaghat, P.O. Waknaghat, Teh.- Kandaghat,
Distt. Solan – 173234 (H.P.) INDIA
Website: www.juit.ac.in
Phone No. (91) 01792-257999
Fax: (91) 01792 245362

DECLARATION

I hereby declare that the work contained in the thesis entitled **“Holistic Multilingual Sentiment Analysis on Reviews in Social Media”**, submitted at Jaypee University of Information Technology, Waknaghat, India, has been done by me under the supervision of Dr. Rajni Mohana. The work has not been submitted to any other organization for any other degree or diploma. I am fully responsible for the contents of my Ph.D. thesis.

Sukhnandan Kaur
Enrolment No. 136215
Department of Computer Science and Engineering
Jaypee University of Information Technology
Waknaghat, Solan – 173234
India

ACKNOWLEDGEMENT

Undertaking this research has been a truly life-changing experience for me and it would not have been possible without the support and guidance that I received from many people. First of all, I offer my heartfelt devotion to the almighty for providing me the opportunity to pursue higher studies and bestowing this sacred blessing on me.

I am indeed indebted to my supervisor Dr. Rajni Mohana, Associate Professor, Department of Computer Science, Jaypee University of Information Technology, for her help and support during my research work. I feel privileged to express my sincere regards and gratitude to my supervisor for her invaluable guidance, constant supervision, fruitful discussions and constant encouragement throughout the course of my research work. The critical comments, rendered by her during the discussions are deeply appreciated.

I pay my deep tributes to all the researchers in the world around, working for the development of Science and Technology for the betterment and enlightenment of the society and feeling to be the part of that community gives me a great pride and pleasure.

The award of the degree Doctor of Philosophy is one of the hardest deserving achievements. People struggle for it and achievement is not easily found. In this regard, I am grateful to the University and express my deep sense of gratitude to its Hon'ble Vice-Chancellor Prof. (Dr.) Vinod Kumar, for providing this great opportunity to me.

I convey my sincere thanks to Prof. (Dr.) Samir Dev Gupta, Director & Academic Head, Prof. (Dr.)S.P. Ghrera, Head, Department of Computer Science and Engineering, Department Research Committee and all other faculty members of department for their motivation and support during the research. I would also like to show my gratitude to the department, for the provision of the best equipment and pleasant office environment required for good quality research. I would also like to thank all administrative, library and supporting staff of the University for providing the comfortable environment and help.

This research work would not have been possible without the cooperation and support extended by fellow research scholars and friends. In this regard, I would particularly like to deeply appreciate the great help of my friends Ramanpreet Kaur, Smita Agrawal, Ambika, Jatinder who always kept the healthy research environment and extended their full cooperation during my research work. They have given me support and a joyful and wonderful university life.

My heartfelt thanks are to my mother Mrs. Paramjeet Kaur and father Mr.Aman Singh Johal, without whom, I would not have had the courage to embark on this journey. Without their support, this research work would not have been possible. They supported me unconditionally throughout my research work. It is difficult to find adequate words to express my appreciation for the help and support given by them.

I wholeheartedly thank my parents, my sisters Gurpreet and Khushkiran for their enduring patience and for providing the moral support during the course of this work.

I wish to extend my gratitude to everyone, who had contributed directly or indirectly to my thesis.

Sukhnandan Kaur

ABSTRACT

Decades ago, people used to represent their opinion by writing it manually or by speaking at public places. These reviews are further taken as an advice for the betterment. To process this data was a complete manual task. As the usage of internet grew people started sharing their views regarding any entity through emails or social platforms. The intensification of data over the social media makes the task of deducing valuable information a bit complex. Automatic deduction of sentiments from web data is considered as a process of sentiment analysis. An algorithm devised for the same is known as sentiment analyser. Use of Sentiment analysers is at the peak for various enterprises to find the loopholes in their product or services. An optimal sentiment analyser is the one which works rationally as humans. The goal is thus to fill the research gaps associated with the effective sentiment processing.

Sentiment analysis integrates many subtasks i.e. Named Entity Extraction, Anaphora resolution, Sentiscore, Feature extraction, etc. Effective pre-processing yields better results for all the natural language processing tasks. The reason for pre-processing of the data is that people use slangs, long tail words, multilingual content and visual language such as emoticons. People use unstructured format of writing along with all the above mentioned categories these days. The process to handle slangs, misspelled words, etc. is called as normalization.

The primary aim of this study is to have effective pre-processing of the content i.e. normalization. Normalization here deals with two aspects: one is to deal with slangs and another is to deal with emoticons. In this study, a technique is used where each emoticon is mapped corresponding to its meaning for generating Sentiscore, instead of just adding or subtracting one for positive and negative smiley respectively. Slangs are also handled effectively by using cross word dictionary and corpus. It is aimed to get better results for pre-processing.

This thesis also puts light on how to deal with multilingualism. These days internet provides the facility to people for writing in any of the desired language or mix of different languages. It makes the task of sentiment analysis more complex. The betterment of sentiment analyser is based on processing this data regardless of the language in which it is written. Multilingualism also comes in the form of macaronic content. In this thesis, a complete analysis of macaronic content is discussed with the proposed technique.

The next objective of the thesis is to present some new results of investigations, demonstrating an application of Temposentiscore for problems related to categorization of reviews in web. Deal with obsolete reviews is the next concern of this study. These reviews may result in biased sentiment analysis which may or may not present the current scenario. To remove this limitation, we are trying to implement temporal sentiment analysis of reviews by providing more weightage to latest reviews. Further, sentiscore is redefined in terms of temposentiscore. Finally, the study mainly emphasizes the need to fulfil the research gap.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SA	Sentiment Analysis
NB	Naive Bayes
SVM	Support Vector Machines
HMM	Hidden Markov Model
PMI	Pointwise Mutual Information
RRM	Robust Risk Minimization
TF	Today for Future
NER	Named Entity Recognisation
PoS	Part of Speech
P	Precision
R	Recall
F	F−measure
FO	Fallout
A	Accuracy
SVO	Subject Verb Object
OVS	Object Verb Subject
NLP	Natural Language Processing
TS	TempoSentiscore

Chapter 1

Introduction

The flow of huge data over several social platforms provides us valuable information. This information is in the form of online journals, remarks, reviews, etc. People these days prefer sharing their perspectives about anything over these social platforms i.e. item, individual or place.

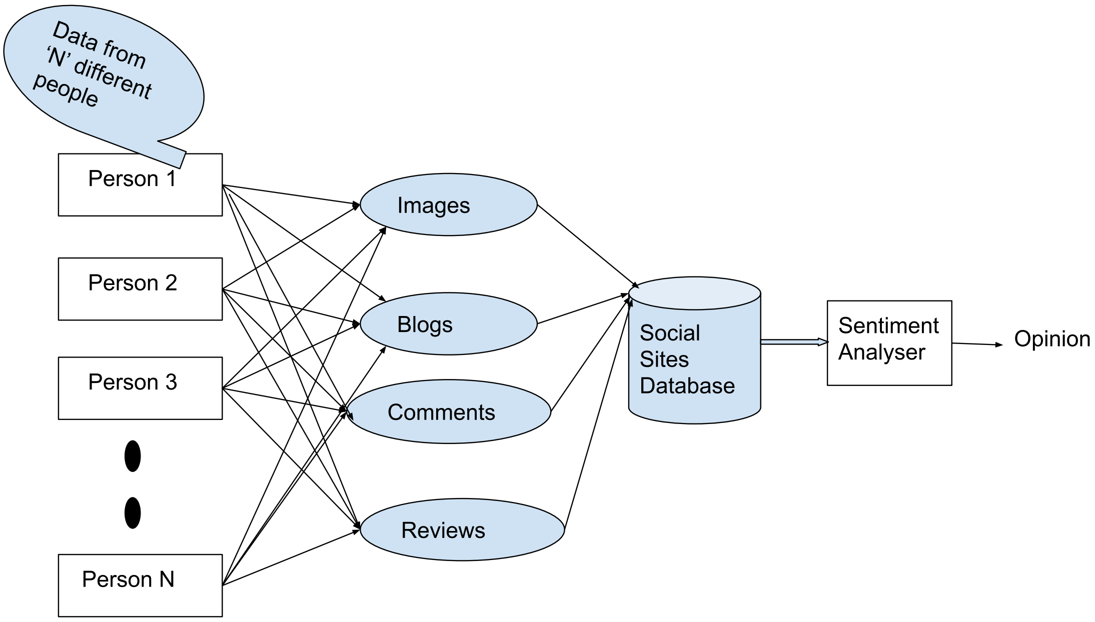


Figure 1.1: Abstract View of Sentiment Analysis

The task of processing such a huge data along with data exhibit in existing studies becomes cumbersome. This tremendous amount of social information opens extensive opportunities for smooth running of advertising endeavors and measures the campaign

Table 1.1: Average Information Shared Per Second [1]

Sr.No.	Social Platform	Rise of Data Per Second	Rank
1	Skype	387,105	3
2	Facebook	39,852,495	1
3	Twitter	19,92,150	2
4	Instagram	172,050	4
5	Reddit	63,045	5

impact on the proficient decision supportive network.

Two decades ago, communal sites which are at the peak these days in terms of usage did not exist. However, at present social sites has overwhelmed by sharing perspectives. The growth and rundown in the usage of various sites which are broadly used by people appeared in Table 1.1[1]. Ranking describes the number of requests for use of different sites. Facebook and Twitter are the platforms which shares significant data, this is deduced from Table 1.1 [1]. Hence, human-oriented extricating information from this enormous data accessible on the internet as appeared in figure 1.1 is taken as a difficult task. It also obstructs the way toward taking the correct decision. The exponential development of data on the websphere quickened the need of removing insignificant data for extracting information. The optional appraisal of individual's experience has done on the basis of resource rare languages and on the resource-rich languages.

Its domain is different from a little individual thing around the world. The need of automation of processing this huge data raises the requirement of improvement in procedures in the field of sentiment generation.

The junction point between online networking and human developed data emerged researchers to work in collaboration in the territory of SA. Sentiment analysis approached us through various circles of our day-by-day lives, regardless of our understanding. It influences our buying habits, work, and so on. Sentiment Analysis is an integral part of decision support system.

1.1 Definition

Sentiment Analysis is a science of reasoning what people think and perceives about any entity. Primarily, the science was used by linguistic experts for various reasons. However, being a dynamic study, Sentiment Analysis has found many more users including businesses and customers. More and more businesses are investing in study of customers sentiments for their products and services. The study allows them to take appropriate actions to enhance their sales in present and future.

Definition 1 [64] : Sentiment analyser consists binary tuples.i.e.,

SA = {e, s}

where,

$'e'$ entity about which the sentiment is generated and $'s'$ is sentiment associated with that document.

Here, the output is considered as the only entity offering an opinion. The process does not take in account as who is giving the opinion. Also, a person can anytime change his/her opinion and thus there will be deviation in the study. This causes a debate and researchers formulated new definition.

Definition 2 [64]: There are four tuples taken into consideration for sentiment analysis.

SA = {a, o, p , t}

where, $'a'$ is the aspect of the any entity,$'o'$ is the sentiment, $'p'$ is the person offering opinion, and $'t'$ is the time when the opinion was given. Here, it was realized that any document can have more than one entity giving opinion. This again instigated the need of another definition.

Definition 3 [64]: SA is now considered as a quintuplet. It consists entity(ey), aspect(a), sentiment(o),opinion holder(p) and time(t). i.e.,

SA = {ey, a, o, p, t}

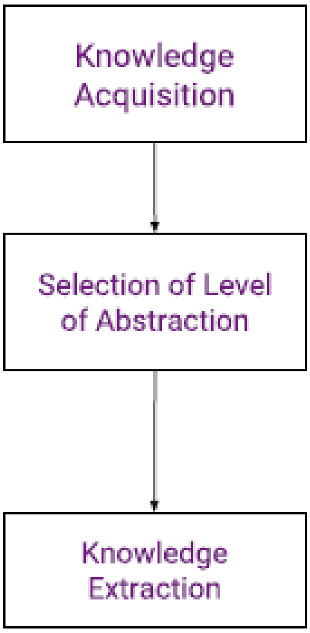


Figure 1.2: Component of Building Sentiment Analyser

1.2 Segments of Sentiment Analyser

Figure 1.2 demonstrates the steps to build a sentiment analyzer. The primary step is knowledge acquisition i.e.to construct a knowledge base. Secondly, focus on identifying the level of abstraction. Secondary step is of deducing the method of extracting or updating the knowledge base.

1.2.1 Knowledge Acquisition

The first step is to train the analyzer to gather more information about the real world entities along with implicit or explicit sentiments. This enables the analyzer makes best use of existing algorithms and extract real-world sentiments. This training is performed by using various sources including experts, lexicons and more. Once collected, the compilation of so many resources is termed as knowledge base.

Knowledge base is the elementary unit of the sentiment analyzer and is made with the contribution of researchers from various fields i.e., literature, knowledge engineering, local people, books, etc. Linguistic reserchers play a vital role in programming of machines and transfer their expertise to the machines for providing accuracy in analysis. Similarly, lexicons make the knowledge base language rich by supplying synonyms, semantic information about different words in diverse languages. Though efficient,the mentioned approaches is tedious and overpriced. Thus, to simplify, people in the field use various autonomous techniques to supply information to the knowledge base. The process is known as automatic growth of the knowledge base.

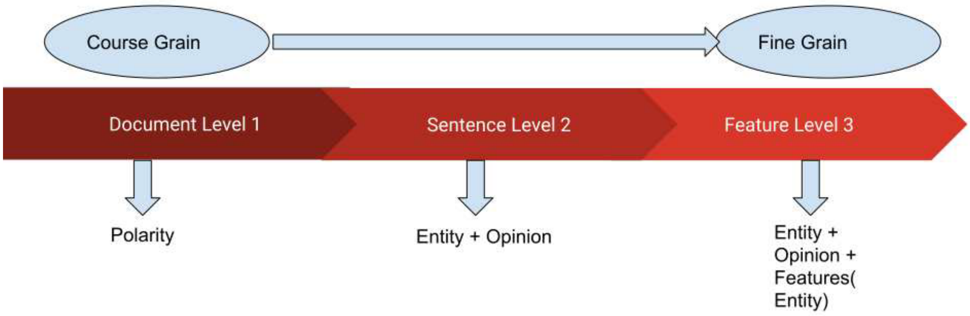


Figure 1.3: Level of Granularity

However, this method is less efficient when compared to the human-initiated processes. Contrary to this, knowledge engineers, do not require expert domain knowledge. The engineers work on the knowledge base alone. They can derive efficient conclusions only if they have a good sense of analysis.

1.2.2 Selection of Level of Abstraction

At the time of inception of sentiment analyser, this was the most important phase of the process. There are three different levels that allow abstraction, namely, document, sentence and feature level shown in figure 1.3. In the document level, the granularity is highly coarse wherein as the level rises, the level of granularity rises simultaneously. The different levels are explained below:

- Document level: It is the uppermost level of the abstraction process. This is where the binary classification is done. The granularity here is obtained at a coarse grain level. It derives the entire document into negative and positive opinions.

- Sentence level: The process of sentiment generation is more or less same in document and sentence level. Extract opinion oriented sentences is the major concern here. This is the primary reason that the level has a higher granularity. Here, the content is are categorized into opinion oriented and non opinion oriented content depending on their nature. It is revealed through in figure 1.4. The analysis here is dependent on the individuals experience of any entity. The basic aim for elimination of non−opinion oriented sentences containing static knowledge and target more on opinion oriented sentences.

  For Instance, 15 languages are widely spoken in my country. Still, local people like to speak in their native language.

  This example has both non−opinion oriented as well as opinion oriented aspect.

- Feature level: This level has a major league in granularity for sentiment analysis. Here, the aspects or features are coded implicitly or explicitly. It further segregated based on domain dependecy. The same is represented in Figure 1.5. This level is used by many researchers to perform studies. Florian et al.[46] repoted four named entity classifiers rule-based, hidden Markov model (HMM), robust risk minimization (RRM) classifier based on regularized winnow method[102] and max entropy classifier. They used these classifiers for English and German languages to deduce named entities. They finally derived that RRM is the best among all. Also, when operated in cross-lingual environment, these methods were not able to derive satisfactory results.

1.2.3 Knowledge Extraction

Various existing models and techniques are used on pre-existing knowledge base and the aim is to derive results. Different models can be used as per the convenience of the analyzer as there is no hard and fast rule.

1. Models: The evolution of machine-learning is related to access the huge compilation of data that may or may not be of great use. The primary aim is to extract meaningful data by sentiment analyser using different machine learning methods. Choosing the right model is critical and at the same time complex. There are no rules for choosing the right model. Basically it is a matter of trial and error, wherein, at times, a wrong model can derive better results than the most suitable method. As per George Box, "Essentially all models are wrong, but some are useful". Models used in the literature are described as follows:

   - Prediction based models: These models analyze the question and interpret the future value. These models also known by future prediction (FP)

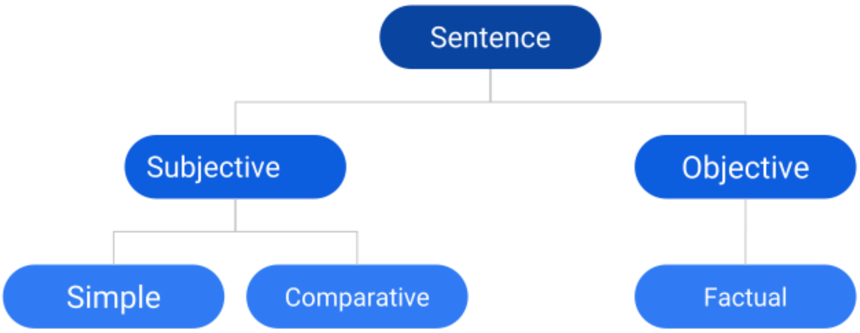


Figure 1.4: Sentence Level

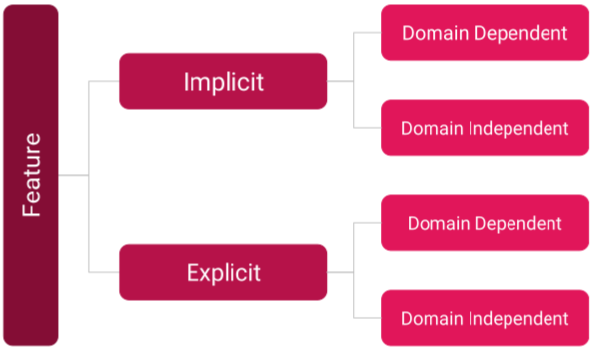


Figure 1.5: Feature Level Abstraction

models. These model's efficiency depends wholly on end-results. Simply put, the accuracy of the predictive values describe the effectiveness of the model. The most commonly used models include Naive Bayes, ARM [80], etc.

- Descriptive models [64]: These model are used for analysis summarization. Various techniques like Classification, Clustering and ARM[80] are used to under these models of SA.

- User-sensitive models [64]: These models of sentiment generation highly useful for interpreting the actual sentiment associated with the review. It takes the user into account while processing.

  For example, a professional camera is well-suited for a photojournalist but cumbersome for a novice user.

- Models based on author authority: These models advocate that the authority of a document often stimulate readers influenc through review [64].

Simply put, a product endorsed by a celebrity is more likely to influence the audience than a product endorsed by a common man.

2. Learning techniques: Number of learning techniques are used for natural language processing like NER, summarization, fake review detection, etc. These learning techniques basically fall in 3 broad categories: supervised , semi-supervised and unsupervised.

    - Supervised learning: When the training data is fully labeled, then supervised techniques are used.
    - Semi-supervised learning: This type of learning is used at the time when training data is partially annotated.
    - Unsupervised learning: It includes the use of algorithm when the data is completely unlabeled.

1.3 Architecture

There are three basic phases that complete the architecture of the sentiment analyzer, as shown in the figure 1.6. The most important part of the sentiment analysis is the pre-processing of data that takes place at first phase. Lexical analysis, morphological analysis, syntactical and semantical analysis are performed at this phase. At second phase, a user is free to use any learning approach for sentiment analyzer i.e. Supervised, Semi-supervised and Unsupervised. Finally, at third phase, the Sentiscore is

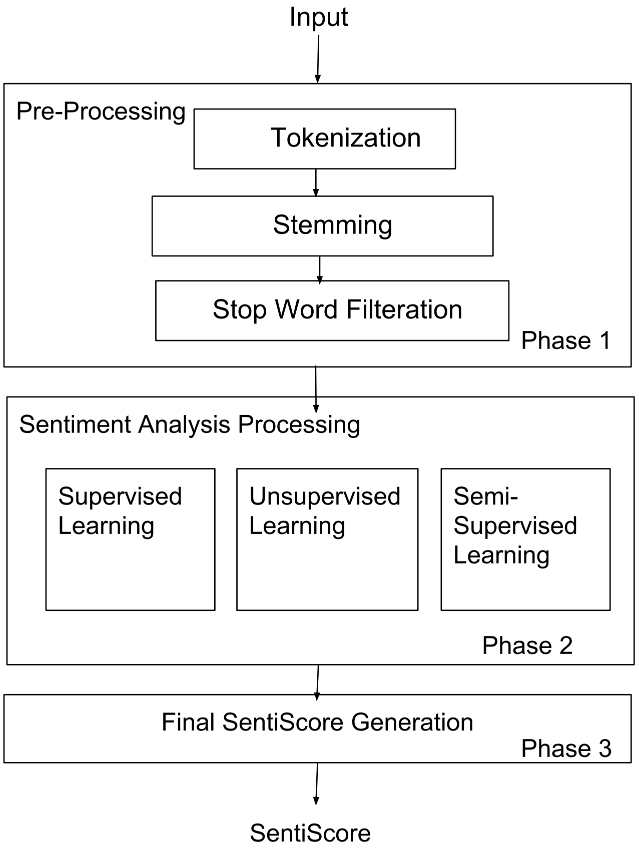


Figure 1.6: Architecture of Sentiment Analyser

calculated.

1.4 Applications

The prime motive of sentiment analysis research is to categorize the content into positive, negative and neutral category. Initially, it was based on a rating system for classifying the data. Later, it shifted from numeric rating to content for classification of reviews or deducing star rating. Recently, the focus is on the prediction of individualś opinion about any entity in various projects like future leader in politics, revenue earning and many more. There are many techniques for sentiment analysis, most of them comprise the use of social data. A few popular applications are explained below:

1. Popularity: The increase in the popularity of an individual entity can be deduced using sentiment analysis. Huge social data reservoirs are used and a combination of text analysis and natural data processing is applied to get the results.

2. Refinement of the product: Through the negative reviews during analysis, manufacturers find out the loopholes in their product and refine them to meet the customers expectation.

3. Mood prediction: Through sentiment analysis, the mood of the audience against any upcoming event, rule, product sale, etc. can be derived.

1.5 Contribution

The contribution of this thesis is in the field of sentiment analysis. It is mainly focused on the extraction of sentiments hidden in the social data present over the internet. The notion of internet makes us to meet with social data and methods. This overcomes the difficulty of gathering opinion from large population manually.

Social data is available in many forms, i.e. online reviews, comments, twitter messages about elections, blogs for any event, or reviews of movies, etc. The primary task of sentiment analysis is classifying the opinion expressed in this social data into positive, negative or neutral.

From past decades, Sentiment analysis has been becoming the hot topic in the research field. For instance, IBM SPSS [2] provides a way to refine product or any service based on user opinion about that product. Also LexisNexis [3] uses news media to analyse consumers content for brand perception. Apart from all these, many other algorithms for sentiment analysis are in trading by various organisations. Motivated from the observations mentioned above, this thesis addresses the following:

The thesis focuses on how to deal with the text containing slangs, emoticons, misspelled words, etc. For effective processing, the need to normalize the text by changing emoticons into senticons, slangs to lexical words in textual data. To examine the effect of normalization over the performance of classifiers. The experiment consists of annotated datasets. This study also considers the fact that each emoticon has its own significance different from others.

This study further explores the task of sentiment analysis from textual data by borrowing the concept of macaronic content. Macaronic content that postulates the indulgence of more than more language in a single document. Manual annotation of the dataset for applying various data driven approaches used in sentiment analysis. Addressing some of the fundamental questions of sentiment analysis, this thesis brings forth the methodologies to carry out multilingual sentiment analysis from the textual data along with macaronic data.

It also proposes an algorithm for temporal sentiment analysis. This makes the final results independent or less dependent on obsolete reviews. Study reveals that proposed system has a potential to expand the sentiment analyserś simplistic temporal view. It introduces a new terminology TempoSentiscore. This captures temporality in sentiment generation using explicit and implicit time variants. To examine its effect over the star rating, empirically derived equation is provided for tempo-sentiscore. It helps in understanding the temporality in terms of generation of sentiscore.

1.6 Organisation of the Thesis

This thesis objective mainly concerned for generation of flexible and reliable natural language processor. The effective preprocessing also increases the reliability of automatic decision support system. It is achieved by considering the temporality of the data implicitly and explicitly. The high level architecture of the system is shown in figure 1.6. This work contains the novelty as it combines natural language processing along with textual analysis i.e. extraction of emotions. Furthermore, analysis of such data through implicit temporality or analysis of informal social data proves to be quite challenging. The thesis is organised as follows:

Chapter 1 of the thesis focuses on basic concept of Sentiment Analysis and how it effectively increases the reliability of the decision support system. It shows the evolution in sentiment analysis through the varying levels of granularity. Applications are also listed in this chapter.

Chapter 2 discusses the background of sentiment analysis. Research gaps show the need to focus more in area of sentiment analysis. Motivation behind the work is also discussed. This helps to generate the prime objectives of the thesis. Objectives are framed at the end of this chapter.

Chapter 3 describes the need to normalize the web content for effective processing. It also strengthens the fact that effective pre-processing gives better results. This chapter primarily focuses on the cleaning of the data at the early stage of sentiment analysis. It increases the reliability of the decision support system.

Chapter 4 of the thesis focuses on how to deal with macaronic content over the web. Along with it, it also describes on how macaronic language is treated differently from multi-linguality. Two languages are taken into account for consideration to deal with macaronic text i.e. Hinglish (Hindi and English).

Chapter 5 discusses the reliability of the decision support system. It shows the need of considering higher granularity level in the perspective of time. Implicit and explicit temporal aspects are discussed in detail. Further, the impact of temporality is also shown over star rating.

Chapter 6 finally concludes the presented work in this thesis. It also highlights the future scope of the work that could be further carry forward in the domain of sentiment analysis.

Chapter 2

Related Study

Decision making capability of a person is highly influenced by the emotions associated with the entity. Conventionally, people used to consult friends and users for their experience about any product while making buying decisions or refer to political forums while deciding whom to vote in the upcoming elections. The suggestions for the same were restricted with the geographical limitations as people′s reach was limited. However, with the advent of internet, the reach has expanded to exponential levels and now, one can not only get hundreds and thousands of reviews but also collect them from across the globe. As we have quoted in chapter 1, a huge pool of people users are active on social media sites. With such a large number of users the internet is overloaded with huge data. This data is very important for sentiment analysis. Traditionally, the primary motive of Sentiment Analysis was to diffuse the sentiments signals into binary forms, i.e. positive and negative. As human experienced the technical hike in machine learning, there was a simultaneous increase in the level of granularity, i.e. rising refinement. In the early 2000s, the primary motive of researchers was to check the document for polarity. This was the case when the analysis was done at document level. However, later it was focused on sentence level (only subjective sentences were considered) and currently, it is more concentrated over entity/feature level.

2.1 Various Stages of Sentiment Analysis

There are three different stages for sentiment analysis as described in chapter 1. The extent of granularity rises with each level. It is finest at the feature level while on document level one can experience the coarsest level. Undermentioned is the work done at each level by different researchers.

1. Document level: This first and foremost stage in abstraction for sentiment analysis. The binary classification of any document is performed at the same level. The extraction here is done at the coarse level of granularity. With sentiment analysis at the document level, negative and positive opinion in the entire document is segregated. Initially, when SA was primarily introduced, document level work was high in demand.

   The majority of work was focused based on the assumption given by Liu et.al. [64] which states that a document is focused on a single entity only. Thus, any work done on a document extracts the opinion about that single entity. However, with more and more people offering different assumptions about an entity, the assumption given by Liu et.at. [64] was disregarded.

   Later, Hatzivassiloglou et.al. [52], and Moghaddam et.al. [73] came up with a theory about SA which was based on a major element of English language resources, adjectives. In literature, it is observed that there are few exceptional

Table 2.1: State of the Art Sentiment Analysis at Document Level

S. No.	Author	Work done	Trade off
1	Turney (2002)	Results are obtained through unsupervised learning approaches	High frequency words were neglected which sometimes contain valuable information
2	Pang et al. (2002)	Several machine learning algorithms were used like NB, SVM, etc. This study found good results in SVM and results are degraded when NB was used.	The performance was not good in aspect oriented SA
3	Liu (2012)	It did not work well in the environment containing opinion from many persons	It was failed when the opinion was from different people regarding the same entity
4	OConnor et al. (2010)	Opinion was extracted from micro blogs	Inappropriate messages selection while sentiment analysis was performed
5	Thelwall et al. (2011)	It has shown the bend towards negative in any event	Different geographical time boundaries hindered the performance of the system.
6	Hung et al. (2012)	SentiWordNet was used along with Cosine similarity for SA	Fake reviews were also considered which gave bad results
7	Baccianella et al. (2010)	SENTIWORDNET 3.0 and WORDNET 3.0 were used to carry out the task of sentiment analysis. 20 % growth was shown using semisupervised learning approach	Specific words have different values when compared with human orinted values. E.g. for the term *'bad'* SentiWordNet gave the results as $pos = 0.625$, $neg = 0.125$, $obj = 0.25$. On contrary, the human generated score was $pos = 0$, $neg = 1$, $obj = 0$, which was conflicting (Brody and Diakopoulos, 2011)
8	Bollegala et al. (2013)	Results were formulated using supervised along with unsupervised learning algorithms in domain independent environment	Word sense disambiguation was not taken care.
9	Maas et al.(2011)	Semantic nature of the content was considered in this study	Performance was degraded for ambiguous words i.e., cold coffee was taken as positive while cold burger was taken as negative
10	Xie and Wang (2014)	Composite words are considered like idiomatic phrases, proverbs, etc. along with simple words for Chinese language	Accuracy was taken as performance metric which obtained as low value.

linguistic rules i.e. *'and'*, *'nor'*, *'but'*, etc. This theory was based on a study done on 21 million words from English literature. However, the algorithm did not consider as efficient because its basic requirement was to have all the training data to be completely labeled, which means that the study did not considered unlabeled data including adjectives, and thus, the algorithm ran down in terms of accuracy. Turney et. al. [95] put forth a theory of unsupervised learning for sentiment analysis. They used a study under which adjectives and adverbs that helped out in calculating the orientation. The algorithm included three steps mentioned below:

Step 1: This step involved extraction of phrases that contained adjectives or adverbs along with a context word that further helped in determining the orientation that used PoS tags[19].

Step 2: Estimation of the orientation of the phrases extracted in the step 1 was performed through the pointwise mutual information (PMI) [94]. Here, the extent of convergence for positive polarity terms and its corresponding negative reference word was the base to compute the opinion orientation of a phrase.

Step 3: The average opinion orientation for all the reviews is computed and if positive, it is considered as recommended. However, the algorithm did not deduce good results when used in processing different languages. This was primarily because of the different linguistic rules of languages.

Pang et al. [80] proposed an entirely different approach for finding the SA. To ensure the classification was done with utmost accuracy, researchers focused to use several supervised learning approaches as mentioned in literature, i.e. NB,SVM and entropy. Their work concluded that machine learning methods were more effective than any human-based method for unigrams and bigrams. These machine learning methods were also more effective in computing topic classification[70] than for SA. The study also concluded that SVM offers more accurate results than Naive Bayes. The entire project was based on polarity checking for positive as well as negative words.

O'Connor et.al. [76] used explicit temporal methodology along with the existing methods. Under his study, they used unsupervised method to determine the political status of a person and included time series to calculate the sentiment. However, the different time frames across the globe gave them a serious problem as it created problem in evaluating the opinion of a document while taking in considerations the online reviews, posts on various social platforms submitted by different humans belong to different zones of the globe. Later, Thelwall et.al. [90] added a three hour burst time to treat the temporality for sentiment analysis and the geographically different time frame problem was also faced by O'Connor et.al. [76].

Larsen et.al. [62] proposed a method wherein they used term frequency count for extracting explicit features. Any term that had a high frequency were considered as candidate feature. Then, all the candidates features were clubbed and the results indicated that continuous center adjustments give better results. They also used adequate mean vector damping technique to adjust the center in clusters but did not considered the extra time. The frequency count approach used here indicated the number of non-frequent items that are usually ignored despite being a valid entity.

Hung et al. [54] also proposed a modification process for the document quality. The method was divided into 5 stages:

- Strong
- Normal
- Weak
- Redundancy check
- Fake data detection

Researchers used cosine similarity between the WOM (word of mouth) in document X and the product description Y, associated with it. They also used base lexicon, i.e. WordNet [72] from different languages and applied various translation or transliteration schemes to make many lexicons and utilize in generalizing the mining of opinions. Sentiscore was assigned to each word from the standard WordNet by Baccianella et. al. [10] and was termed as word score. This score was calculated through an averaging method of all scores including negatives and positives for each word present in a text span that has a relationship with feature M. The same was done by using equation 2.1.1.

$$WordScore(W) = \frac{1}{q}\sum_{t=0}^{q} posScore(t) + \frac{1}{q}\{-\sum_{t=0}^{q} negScore(t)\} \tag{2.1.1}$$

where,

posScore(t): positive score obtained for synset t.

negScore(t): negative score obtained for synset t.

q : count of all the synsets.

The equation helped in determining a uniform number for every term which is termed as SentiWordNet [43]. It was observed that a few researchers have already performed inter-domain analysis. Bollegala et al. [18] came up with a theory with use of sentiment analysis in inter-domain analysis. They used both rule based and machine learning approaches for their work. The performance of the method faced no competition from the existing baseline methods. As a result, SA was focused on the meaning of a word instead of focusing on string matching methodology. Maas et al. [68] devised an algorithm for finding the similarity among different words. The theory has combined both supervised and unsupervised learning techniques. The vector-based model that they proposed deal with semantic and sentiment similarities between different words. The theory failed to lead in domain dependent sentiment analysis. It was believed that each language has a huge reservoir which contains simple words as well as complex words such as idioms, phrases, proverbs and more. These words are also important for sentiment analysis. Xie et.al. [100] proposed an unsupervised

Table 2.2: State of the Art Sentiment Analysis at Sentence Level

S. No.	Author	Work done	Trade off
1	Pang and lee (2004)	Differentiate the sentences into opinion oriented and opinion lacking segments. Afterwards, opinion lacking sentences are removed as these did not exhibit any opinion expression. Supervised learning approaches are used for the performance evaluation. Semantic orientation of opinion bearing terms are also considered. Due to independence of different attributes Naive Bayes gave better results.	This study did not consider Word sense Disambiguation which hindered the performance somehow.
2	Hatzivassiloglou and Wiebe (2000)	Adjectives were taken into account for the classification of reviews	This technique gave bad results for the languages which did not get appropriate consideration in the literature or having no or less resources.
3	Boiy and Moens (2009)	Multilingual content was considered including Dutch and French along with universal language English. Unstructured content was also the focus of this study.	Did not work well for misspelled words.
4	Wilson et al. (2005)	Obtaining prior polarity got lowest priority as compared to polarity based on the context or domain.	Binary classification was considered instead of positive, negative or neutral.
5	An and Hagiwara (2014)	Emotion was deduced based on top five sentiment bearing words. i.e., adjectives	Failed to perform well for informal or unnormalized content.
6	Liu (2012)	Some words may have different semantic nature depending on the domain in which it is used. This was considered in this study	Informal text processing was untouched.

technique that extract sentiment from the annotated Chinese reservoir enrich with idioms. They noticed that this theory worked well with unannotated dataset. Sentiment analysis work on document level is recorded in table 2.1. It indicates the importance of segregating fake reviews from authentic reviews for quality results. The study also deduced that most of the times term frequency method was the best for named entity extraction. At the same time, term frequency method also had a shortcoming as it often neglected words that may carry some important information. It was also deduced that SVM should be used for complex sentences while Naive Bayes was considered as an efficient method for simple and short sentences.

2. Sentence Level: This level is more or less same with the level described in previous section for sentient generation. At sentence level, the primary focus of SA is to detect the subjectivity, because of this the granularity at the document level rises. In a document, there are sentences that can be categorized as subjective or objective in nature as described in chapter 1. Here, for analysis the opinion depends upon the experienced of an individual regarding the entity. At this level, the primary focus of the researcher is on the subjective sentences and thus, all the objective sentences are ignored.

   For instance, 15 languages are widely spoken in my country. Still, local people like to speak in their native language.

   Above Example is the combination of subjective and objective sentence.

Hatzivassiloglou et.al.[53] studied how the orientation of adjectives in SA can affect the subjectivity at the sentence level. This study was increased the performance level of sentiment analyser.The main focus here was to find out whether a word was subjective or objective depending upon the presence of the adjective in that particular sentence. The method failed in cross-domain, because a word that had a positive value in one domain might had a negative value in the other. To ensure performance enhancement, the researchers used a huge tarining data consisting adjectives, which was a costly affair. Pang et.al. [78] also stated that compression does not affect the overall polarity. Compression here means excluding the very part that isn't relevant or has no contribution towards generating opinion. Several machine learning approaches were used for SA at sentence level. Graph based minimum cut detection method was used for detecting the subjectivity.

However, adaptive methods has shown lower accuracy because different words have different sentiments associated in various languages. Because of this disambiguation of polarity, it becomes harder to find the polarity of any sentence.

For instance:

National trust helped many poor people.

There are two words $'poor'$ and $'trust'$ in the above sentence, these do not consider as opinion bearing word. So, there is no use for generating opinion based on these words. Any meaning can be extracted only if they are considered as neutral which is again a general rule of SA. Wilson et al. [98] used subjectivity clues to find out the polarity of the context. The process had two-step as explained underneath:

Step 1: Clues classification as polar or neutral. A total of 28 featured were extracted which were summarized as: Word features consists of context, prior polarity, etc. Modification based linguistic features including intensification, proceeded by adjective, dependency phrase info, etc., Sentence based features to find pronoun in sentences, etc. and Document based features including topic of document.

Step 2: All the clues instance that were categorized as polar were further segregated based on the nature of their polarity. Here, they used 10 different features. Broadly summarized into two categories: term features like tokenization, word polarity identification, Polarity features consists of negation, modified polarity, etc. From their work, they evaluated the fact that aggregation of these features helped in obtaining better results. Liu et.al. [64] addressed two aspects which hinders the performance concerning sentiment analysis. First, context dependency was considered for semantic orientation. Second, sentences may have several opinion oriented terms. They suggested a key for the problem and aggregated several opinion oriented terms exist in the same sentence. Earlier, whole work was concentrated on a universally accepted language i.e., English because there were lesser techniques available for translation and transliteration. With better translation techniques, SA became more advanced. At presents, researchers are focused on conducting SA on as many languages as possible. Boiy et.al. [17] came up with a new technique of extracting opinion from several social platforms containing reviews, posts from various users, etc. in different languages, i.e. English, Dutch and French using number of supervised approaches. An et.al. [8] proposed a method of deducing a persons emotion on the basis of top five adjectives. Depending upon the adjectives a the mood of person or nature can be decided.

For example, For the first 5 positive adjectives, it was considered that the person was in a positive mood otherwise not.

The complete work is summarized in the table 2.2 and suggests that it is important to have analyzers work well even for unstructured data. Though multilinguistics has been already addressed but still there is a lot to consider. It is important to carry out such researches in resource scare language. It is also important to handle ambiguous words with utmost care for effective sentiment analysis.

3. Feature level: This is where the granularity is at the highest level as discussed in chapter 1. It shows the classification of implicit features along with considering features explicitly which can further be classified as domain dependent and independent. There is a huge work submitted by researchers at the feature

level. Florian et.al.[46] has proposed four NER classifier's rule which was focues on hidden Markov model, robust risk minimization classifier based on regularized winnow method [102] and max entropy classifier to extract different names entity in two languages, i.e. English and German. The study concluded that RRM method is better than all other methods. The study also suggested that the results extracted through this study were not efficient enough when done in a cross-lingual environment.

Zhu et. al.[105] used ARM (associated rule mining) method for determining frequent features in Chinese language. They also used Apriori algorithm [7] for calculating the frequent items. This approach was used because it was considered that it will make the process swifter than other processes. Later, topic co-relation filtration method was used for extracting candidates features out of all the calculated items. The only shortcoming of this method was that all the non-frequent items were neglected in the process. Yi et.al.[101] made use of the keyword approach. Under this approach, a spotter is used to extract the feature after explicitly mentioning any predefined arbitrary item set. The method collects all the words under one topic as per the synonyms found. They also used disambiguation for finding like items that are related to any topic.

For example, a pre-defined items set includes Sun Microsoft. During the process of tokenization, there are possibilities to differentiate Sun and Microsoft. He also proposed that there should be a association between the two entities to extract them. This was executed by using word frequency-inverse document frequency count that relies on off-topic and on-topic items[101].

Kucuktunc et al.[61] also worked on supervised learning approaches and derived various features of SA including temporality in sentiments, calculated the impact on sentiment geographically and contextual dependency. The primary finding[61] stated that sentiment analysis was strongly dependent on the topic demographic factors. High influence of temporality based on different aspects of time was shown. Zhao et.al. [103] proposed a method that focused on the association of sentiment bearing term and the feature word for extracting the aspect. It was furnished by deriving multi-aspects [104] SA. The primary objective was to ensure that the opinion mining domain was independent. They used PoS taggers to determine association level. At the later stage, researchers were more focused on implicit feature extraction studies than on explicit. Srivastava et.al.[88] used binary grammatical dependencies between different opinion words and feature through PoS tagging in a particular domain oriented environment i.e.,product reviews and studied implicit feature extraction. Ding et. al.[40] stated that contextual dependencies is a part of SA and also suggested how to efficiently handle these context-dependent opinion words. They made intra-conjunction rules under which opinions from both the sides must be considered and had similar polarity. Xie et.al.[100] made use of a basic classifier in order to minimize the dependability of domain by considering phrasal structures

Table 2.3: State of the Art Sentiment Analysis at Feature Level

S. No.	Author	Work done	Trade off
1	Florian et al. (2003)	Named entity was considered for processing. Several machine learning algorithms mentioned in the literature were used along with feature selection. Out of these algorithms, RRM gave good results.	Multilingualism was not considered. Results found for Dutch were not as par as English.
2	Kucuktunc et al. (2012)	Impact of external factors such as Gender, Qualification, Event name and time was considered for sentiment analysis	Explicitly mentioned factors were undoubtedly considered. On the other hand, implicit factors were totally ignored.
3	Zhao and Zhou (2009)	Sentiment analysis was considered in purely domain independent environment	Efficiency was wholly dependent on how effective the trainig data is.
4	Srivastava et al. (2010)	Sentiment analysis was carried out in association with grammatical structure	Domain dependency and independency was the concern. Domain dependency did not perform at par with doamin independent data processing.
5	Ding et al. (2008)	Opinion orientation of several terms or words depends on the context in which it is used. This aspect was considered in this study.	Variation in the structural format of various languages effected the performance of the system
6	Xie and Wang (2014)	Sentiment analysis was done using Chinese idiomatic language resources for specific domain. This gave better results but for the considered domain only.	In terms of Bag of Words only Bigrams were considered.
7	Zhu et al. (2009)	To speed up the processing for sentiment analysis Apriori approach was used. Chinese content was taken for processing	Apart from chinese language other languages were not considered.
8	Yi and Niblack (2005)	Feature extraction was explicitly carried out using keyword approach. The matter of word sense disambiguation was also touched in processing.	Features were not extracted implicitly.
9	Ding et al. (2009)	Features were extracted implicitly and explicitly	System works inefficiently in domain dependent environment.
10	Pontiki et al. (2014)	Higher level of granularity was considered. Therefore, extract named entities along with different aspects	Domain dependency was the reason in lowering the efficiency of the system.
11	Che et al. (2015)	Focus on the importance of preserving the polarity even after removing the redundancy	Semantic orientation may sometimes reduce the efficiency of the system.

that had the similar meaning on a universal ground. Ding et.al.[41] studied the extraction of implicit and explicit entity. They used two different methods for extracting explicit and implicit entity. Pontiki et.al.[82] aimed on developing aspect-based SA. The entire study was based on extracting the entities, based on the aspects and polarity associated with the aspect. By summarizing the SA at feature level as shown in table 2.3, it is concluded that there is very less work done for extracting the implicit feature in a multi-lingual data reservoir. There is a huge scope of work in cross-domain feature extraction also.

2.2 Methods of Sentiment Analysis

To ensure that the huge repository of data is accessed that can either be of great use of completely worthless, it was experienced that machine learning is important. There exist many machine-learning methods that ensure meaningful data extraction through SA. The choice of the right method is considered as necessary as well as complex. There is no framework which can help to pick any model from the choice. Many times a wrong model can offer the right prediction and can be treated as correct as described in chapter 1.

Learning Techniques

Below mentioned are the approaches in literature that adopted for training the system:

1. Supervised Learning: This is followed for system training when the content is labeled. SVM, HMM, etc. are a few supervised algorithms that are used for

this type of learning approach. Pang et. al.[80] suggested using Naive Bayes, maximum entropy and SVM approach for SA for classifying movie reviews into positive opinion and negative opinions. Feature are extracted in this method by performing through the combination of entities that have effective names in the context. The study concluded that the used algorithms did not performed well for sentiment classification as it did for text classification. Florian et. al.[46] also used HMM, a RRM classifier based on regularized winnow methods for named entity extraction along with the existing models used by Pang et. al.[80]. They derived that RRM classifier was the best technique for feature extraction among all. Correlation method for feature extraction was also used after considering the association between bigrams, trigrams and Ngrams in a topic along with a distance measure proposed by Liu et.al.[65]. In a particular domain oriented environment, supervised techniques offered the best results.

2. Semi-supervised Learning: This algorithm offers the best results when the input has a combination of both labeled and unlabeled data. Pang et.al.[78] proposed Graph-based semi-supervised-learning methods based on minimum cuts. To minimize the manual labeling of input data, Etzioni et. al.[44] gave a bootstrapping method. However, efficiency of bootstrapping is widely dependent on seeding to uplift the performance for extraction during the time of training the system. Riloff et.al.[85] also used a semi-supervised technique that means bootstrapping on annotated data through linguistic clues and finding out the

Table 2.4: Confusion Metric

	Machine says YES	Machine says NO
Human says YES	t$_p$	f$_n$
Human says NO	f$_p$	t$_n$

patterns for subjectivity. Bootstrapping method is considered extensive broadcast method used for annotating words, expressions, etc. that has either polarity or subjectivity.

3. Unsupervised Learning: This algorithm is used when analyzing unlabeled or annotated data. There are different rules are made for system training with the use of unlabeled data. Florian et.al.[46] suggested an agglomerative classifier. Later, was adopted for classifying based of their active features and the combinations. Chamlertwat et.al.[23] used lexicon based unsupervised learning algorithm. The technique is termed very useful but less accurate because of the high deposits of annotated data over the web.

2.3 Performance Metrics

Below mentioned performance metrics are used for various natural languages processing tasks including sentiment analysis to analyze the results. The list includes Precision, Recall, F-measure and Accuracy. These measures can be calculated using confusion metric as given in table 2.4 Precision: It is defined as fraction of retrieved documents that are relevant to total detected documents. It is calculated using equation 2.3.2.

$$\left[P = \frac{\substack{\text{number of correct positive or negative} \\ \text{documents detected by the system}}}{\substack{\text{no. of positive/negative documents} \\ \text{detected by the system}}}\right] \tag{2.3.2}$$

Recall: It is defined as the proportion of relevant documents that are retrieved and total number of documents. It is calculated using equation 2.3.3 .

$$\left[R = \frac{\substack{\text{number of positive or negative} \\ \text{documents detected by the system}}}{\substack{\text{no. of positive/negative documents} \\ \text{present in the Gold Standard test set}}}\right] \tag{2.3.3}$$

F-measure: It is a harmonic mean of Precision and Recall. F-measure with $\alpha = 0.5$, means taking Precision and Recall at equal weightage. It is calculated using equation 2.3.4.

$$F = \frac{(\alpha^2 + 1) \times P \times R}{\alpha^2 (P + R)} \tag{2.3.4}$$

Accuracy: it is the fraction of classifications that is correct. It is calculated using equation 2.3.5.

$$A = \frac{t_p + t_n}{t_p + t_n + f_n + f_p} \tag{2.3.5}$$

Fall-out: It is a measure of the proportion of mistakenly selected non-targeted items. It is calculated using equation 2.3.6

$$FO = \frac{f_p}{t_n + f_p} \tag{2.3.6}$$

2.4 Research Gaps

In Future, SA will be used to design and develop a system that has common sense perform at par with human being. This is a major shortcoming in the current state of the art sentiment analysers. Figure 2.1 shows the research gap. Many machine learning methods have been honed for increasing the performance of sentiment analyser. These techniques are also used for transforming informal content to formal content after applying a number of learning approaches as mentioned in chapter 1. The gaps in the research is thus concluded to design a theory of commonsense for natural text analysis.

2.4.1 Current Research Directions

For a sentiment analyzer to become ideal, it should have intellectual capabilities like humans. This can be done by filling the research gap. Filling research gap is important and can be achieved by working on several research directions. Under this part, the ongoing research directions and other features are identified. It demands researchers attention in order to work more accurately to derive an ideal sentiment analyzer.

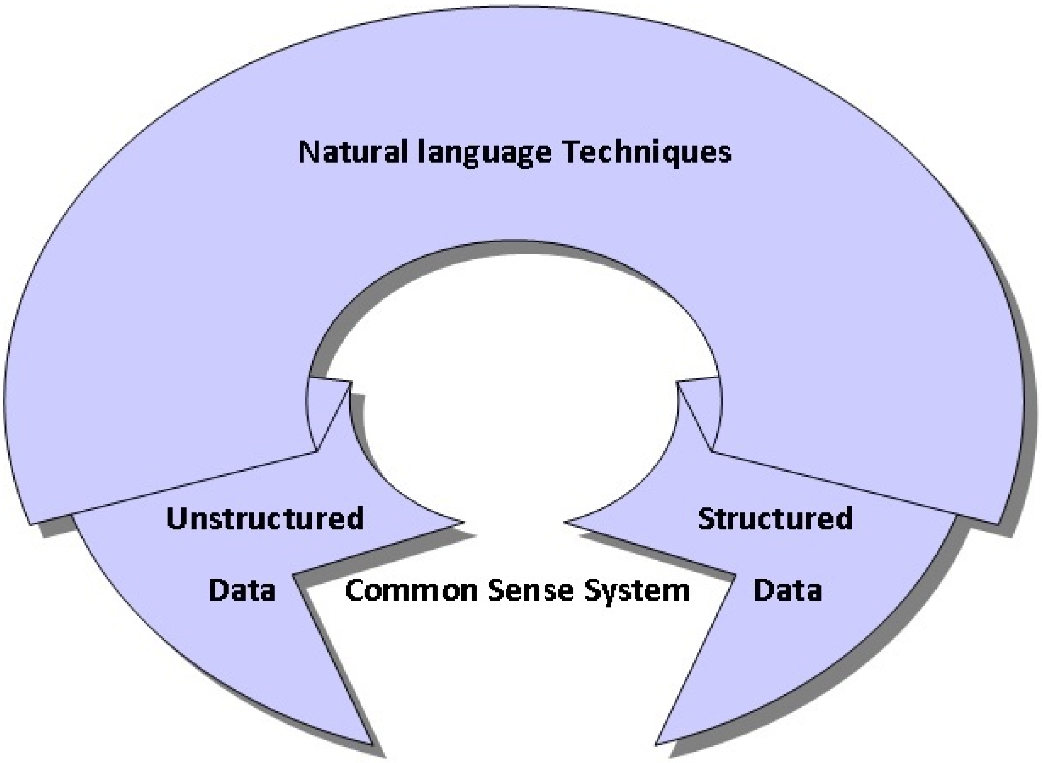


Figure 2.1: Breakthrough of Sentiment Analysis

Fake review detection

There is abundance of electronic data and thus, arranging data for studies is no more an issue. However, there is also an increase in the spam content. The quality of SA depends upon the authenticity of the opinion feeding system in order to extract the sentiments. Fake review detection is a process that allows the identification of people hired by a company to boost their reputation or lower downs the competitors reputation by posting fake reviews. For finding the justified opinion for any product or service, it is important to run a fake review detection. Researchers are more active than ever for identifying useful review amid the bogus reviews and comments for fraud

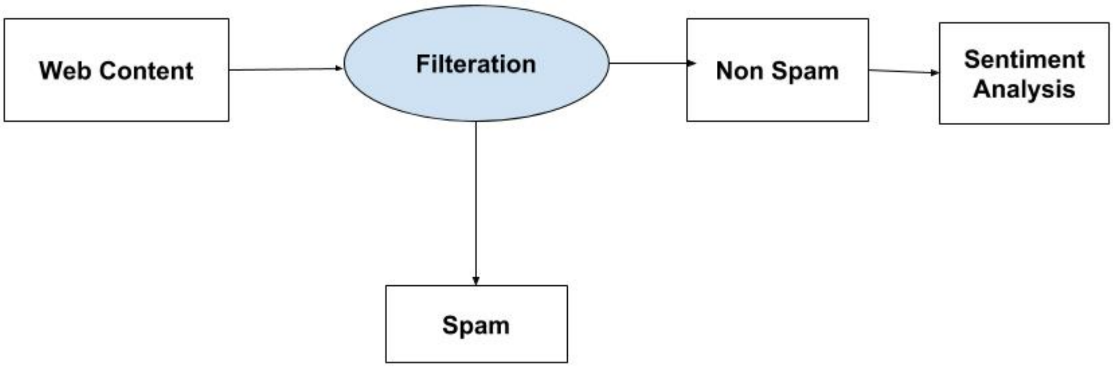


Figure 2.2: Fake Review Detection

Table 2.5: State of the Art Fake Review Detection

S. No.	Author	Group Spammers	Individual Spammers	Domain independency	Content based	Meta Data	Factual information	Results
1	Jindal et.al.(2007)	×	√	×	√	√	×	Accuracy = 0.78
2	Lim et. al.(2010)	√	×	×	√	√	×	H.A. = 0.64
3	McCord et.al.(2011)	×	√	×	×	√	×	A = 0.957
4	Mukherjee et.al.(2012)	√	×	×	×	√	×	H.A. = 0.79
5	Wang et. al.(2012)	×	√	×	×	√	×	H.A. = 0.60
6	Mukherjee et. al.(2013)	×	√	×	√	√	×	H.A. = 0.76

detection [55][96][71] and also to ensure better customer profiling as shown in Figure 2.2. Wang et.al.[97] also considered a few spam reviews in his analysis instead of neglecting them completely. Nitin et.al.[55] categorized fake review detection into two categories-link spam and content spam. The research issues in this process includes:

1. Distinguish the group of spammers or a spammer: It has been understood that more quality reviews are responsible for the betterment of the system

and accurate output for any decision support system that can be expected. This enhances the need of identifying fake reviews for quality analysis. Lim et. al.[63], Mukherjee et.al.[74] and many other researchers worked on the process of identification of group spammers and achieved results at par satisfaction. However, this should not underrate the importance of finding individual review, which is also considered as one of the most complicated and critical task of review detection. Blocking the ids of rubbish reviewers, impose a penalty on them. Few suggestive steps that can be taken to ensure quality of reviews.

2. Domain independence: Movie, restaurant, products, etc. are a few example of domains that demand spam detection. Here, the primary motive of the researchers is to develop a system that facilitates spam detection regardless of the choice of domain i.e. domain independency.

3. Content-based/metadata-based/factual data: Here, three different approaches are responsible for fake review detection as described as:

   - Contentextual approach associated with the data shared by the person in context .

   - Every post has Metadata attached by the server, i.e., time of post, date of post,etc.

   - Factual information is about the facts or figures. Sentences holding factual information are also known as objectivity. Sentiment or emotion about any entity also stimulates its behaviour because of the change in some of the

Table 2.6: Temporality in Sentiment Analysis

S. No.	Author	Explicitly mention of topic keyword	Implicitly deduce topic keyword	Handling the geographical dispersion of time	Forecast analysis	Weightage hinged to reviews w.r.t. time	Results
1	OConnor et.al.(2010)	√	×	×	√	×	R= 63.5
2	Thelwall et.al. (2011)	√	×	√	×	×	P = 0.013
3	Razavi et.al.(2013)	√	×	×	×	×	H.A. = 0.69
4	Dias et.al.(2014)	√	√	×	×	×	P = 0.078
5	Fukuhara et. al. (2007)	√	×	×	×	×	R = 0.78

factual information, i.e. price, income, etc.

Table 2.5 indicates that researchers have not yet tried to include factual information in a SA task or fake review detection. As factual information will be added, spam detection process can be intensified.

Temporal nature of sentiment generation

Temporality is another important dimension that is overlooked by researchers. Researchers usually focus on identifying the opinions. Apart from this, it detects the time when a person offered them, in order to find out the shifts in attitudes over time. There are various applications where temporal aspect plays a significant role including reviewing the success of marketing campaigns, calculating the damage to a brand and also in controlling it by responding to the problem as soon as possible. This means, there should be a system that has capability to assign high weight to recently posted review over any blog or review that was posted in the past.

Below mentioned are aspects where temporal nature of reviews is included in SA:

1. Explicity/implicity in topic word: Many times topic words are deduced explicitly or implicitly. It is suggested that implicit extraction is the best method but at the same time is quite complex.

2. Geographical dispersion of time: When temporal aspect is considered for SA, it becomes more complex as there are different time zones across the globe. To derive a collaborative real-time sentiment for the entire world is thus a very complex task.

3. Forecast analysis: There is active work being done to increase the measures of forecast analysis by taking time into account for SA.

4. Assign weight to reviews depending on time: With changing time, the sense of submitted reviews degrades and the current review becomes more important. It is suggested that there should be active use of hinged weight according to temporality for obtaining time-oriented review.

   Table 2.6 shows how researchers did not take in account any figurative value. This is the reason that less importance was paid to deduce the sentiments based on the standard time scale with the use of meta-data.

Table 2.7: Multilingual Sentiment Analysis

S. No.	Author	Implicit feature extraction	Explicit feature extraction	Translation or transliteration	Automatic expansion of bilingual lexicon	Collocation	Multi-linguality in single Sentence	Results
1	Denecke (2008)	×	√	√	×	×	×	A = 0.58-0.66
2	Banea et al. (2011)	×	√	√	×	×	×	R= 83.15 P = 67.76, A= 69.44
3	Lin et al.(2012)	×	√	√	√	√	×	A = 0.706
4	Hogenboom et al.(2014)	√	√	√	√	×	×	P = 0.062
5	Boiy and Moens et. al. (2009)	×	√	√	×	×	×	A= 0.87

Multi-linguality

The internet is the most advanced and highly utilized medium for communication. To ensure that more and more people reads the submitted text, a writer often chooses English as the medium considering that majority of its readers have the same native language. As an alternative, the writer can also write the reviews in English as well as in his native language. This in turn increases the efforts needed to write a document but also for maintaining it. Banea et. al.[13] found that only 39.4% of internet users know English language. To communicate over the internet, rest of the people use either their native language or any other language that has more supporters in the existing systems. Different aspects of handling multilingualism in textual data are mentioned below:

1. Use of translation/transliteration: There was very less work done on multilingual data, previously. Because there were not quality translation techniquesavailable, majority of the SA tasks were performed on manually built bilingual lexicons. With the advanced techniques of supervised translation and transliteration, the research in SA was enhanced significantly.

2. Expansion of self regulating thesaurus for multiple languages: To enable working in different languages a need for parallel lexicons was experienced. It was practically impossible to have a dictionary that has all the words due to use of new words in natural language. Thus, to work with multi-lingual data, it was mandatory to have an extension of the current lexicon.

3. Collocation: A linguistic term can be translated in a number of ways in other languages. It makes finding bilingual collocation correspondence very hard. The collocation method includes translating a word depending on its semantic nature.

4. Sentiment analysis depending on implicit and explicit nature: The use of translation approaches made it simple for extraction of entities explicitly along with sentiment. However, finding out implicit entities still remained a complex task. The need of a parallel corpus was experienced to perform SA implicitly.

5. Multilingulism associated with sentences: Any sentence that comprises the base language is easier to deal with. However, in a sentence that has multi-lingual

words, most of the foreign language words are considered as stop words which is often a valuable task for extracting opinion. All the present approaches for translation and transliteration are not capable of detecting words without an explicitly knowledge of their language.

Table 2.7 indicates the importance of focusing on multilingualism in sentiment analysis for enhancement of quality.

Unstructured sentences

At present, majority of self driven supervised approaches are capable to work for formal content. Most of the times, structural analysis covers data with a highest degree of organization, straightforward search engine approaches whereas, informal data is essentially the opposite. However, because of no formalism for the same, compiling data becomes more energy and time-consuming task. There is abundance of Unstructured data, which isn't easy for machine to read. Unstructured data is comprised of:

1. Informal Text: Here, no grammatical rules of English language are followed for sentence creation. The structured sentence in English language contains Subject Verb Object (SVO) or ObjectVerbSubject (OVS) format.

   Structured sentences: I am glad to see you. Previously, i was missing you badly.

   Unstructured sentences: glad to see you. Missing you.

Table 2.8: Normalization for Sentiment Analysis

S. No.	Author	Short Words	Normalization based on long tail words	Stemming	Idiomatic handling	Abbreviation	Spelling Correction	Impact of long tail words in SentiScore Generation	Results
1	UzZaman et.al. (2005)	√	√	×	×	√	√	×	-
2	Willett et.al. (2006)	×	×	√	×	×	×	×	-
3	Whittle et.al. (2010)	×	√	×	×	×	×	×	A = 0.58
4	Brody et.al. (2011)	√	√	×	×	×	×	×	P = 0.676
5	Eisenstein et.al. (2013)	√	√	×	×	×	×	×	-
6	Liu et.al. (2012)	√	√	×	×	×	√	×	-

2. Use of slangs (short words/abbreviation/spelling correction): All the internet users enjoy a privilege of the ability of writing the reviews in any desired format. Most of the times, people use alphanumeric words to minimize the number of words to be written.

   Example: It is a gr8 news. gr8 is used instead of great.

3. Normalization(Removal of Noisy Words[20]) and smileys/elongation of words): Most of the social sites allow users to use an image in order to express their mood such as happy, sad or angry. Along with these, people also use different written words to express their extreme feelings such as grttttttttttt in place of

great and lovelyyyyyyyyyy in place of lovely.

4. Stemming: This is a process of finding a root word often termed as morphological analysis.

   For Instance: cat is the base word for cats.

5. Change in strength of opinion: The existence of long-tail terms give either rise or drop in the intensification of sentiments or feelings about any entity. After normalisation, it often loses its intensity which hinders the actual opinion.

The addition of long tail words often increases or decreases the feelings about any entity. Normalization often makes it to lose its intensity which hinders the quality. **For example,** Sorrrryyyyyyy should be generalized as very sorry and not sorry. On concluding remarks for table 2.8, it is observed that sentiment analysis still lacking to consider idiomatic phrases which are used in the textual content and the intensity change due to the simple and long-tail words into actual processing. Researchers need to enquire for this type of content for the rise in performance of sentiment analysis.

2.5 Objectives

The aim of this research is to bridge the gap in the automation of sentiment analysis and decision support system. The objectives are summarised as the following:

1. To normalize web content containing slangs, emoticons and misspelled words.

2. To handle Macaronic language content for Sentiment analysis.

3. To capture the essence of hidden temporality in sentiment analysis.

4. To analyse various supervised learning approaches in different datasets.

2.6 Summary

This chapter introduced the investigation from numerous researchers. It shows the research gap associated with current scenario needs to be tackled. This chapter specified the primary components of sentiment analysers along with granularity levels. The attention has paid to have sentiment analysers work at par with humans i.e., introducing commonsense features to it. Multilinguality, spam detection and temporality are considered as the major concern of this study. The need to pay attention in considering factual content is also highlighted. The concern associated with multilingual data containing several language within a single document instead of one is also revealed. This study emphasizes the need of considering all the mentioned aspects in this chapter. It helps to excel the results of sentiment analyser.

In the forthcoming chapter 3, 4 and 5, the above mentioned issues are discussed in detail. The next chapter explore the effectiveness of pre-processing of data during Sentiment Analysis.

Chapter 3

Proposed Preprocessing Method for Handling Slangs and Emoticons

We begin our exploration by preprocessing, the most important source of text is undoubtedly the Web. The web content is full of unstructured content. People use to write misspelled words, short words due to word limit for various social platforms and the rise in visual language these days i.e, emoticons. It hinders the performance of decision support system by not capturing the exact semantic nature of the content. To enhance the performance of decision support system, it is very much required to process data efficiently i.e. semantically correct. The focus of work is to process the semantically correct and methodologically useful content for sentiment analysis. To find the significant meaning or the replacements of each and every slag is the key concern of the work presented in this chapter.This can be applied to the preprocessing of any textual data for language processing task. This helps in enhancing the performance of automatic decision support system. Various natural language processing (NLP) tasks are carried out to feed into computerized decision support systems. Among these, sentiment analysis is gaining more attention. In this chapter, we propose a novel method of normalization of web data during preprocessing phase.

3.1 Introduction

Natural language processing is a field of computational linguistics and artificial intelligence. It is the key to unlock various decisions using narrative web content. The automation of decision support system widely relies over the performance of natural language processors. Data available over the web sphere in various forms such as text, audio, video or pictures. Due to the arbitrary nature of the language, this data is unstructured in nature. Efficiency of decision support system also gets affected by this unstructured data processing. This may sometimes hinder the performance of sentiment analyser. Thus, affecting the decision support system. As shown in Figure 3.1, initially, data is collected from the various social sites for automation of the decision support systems. Then data is pre-processed to get it in structured format which includes removing the redundant content, cleaning and normalization. Later, various language processing tasks are carried out. Depending on the requirement, the results of the language processor are filtered out for the automation of decision support system. In our work, we have taken the result of sentiment analyzer (SA) into account. The proliferation of web data primarily as communication medium give rise to the existence of unstructured content in the form of posts, blogs, reviews, etc. This web data is rich indicator of peoples reaction for any entity. This reaction of people is analysed and termed as sentiment analysis in the field of natural language processing. Classification of this web data into predefined categories , i.e. positive, negative or neutral is the task of sentiment analyser. The web content is usually the raw data

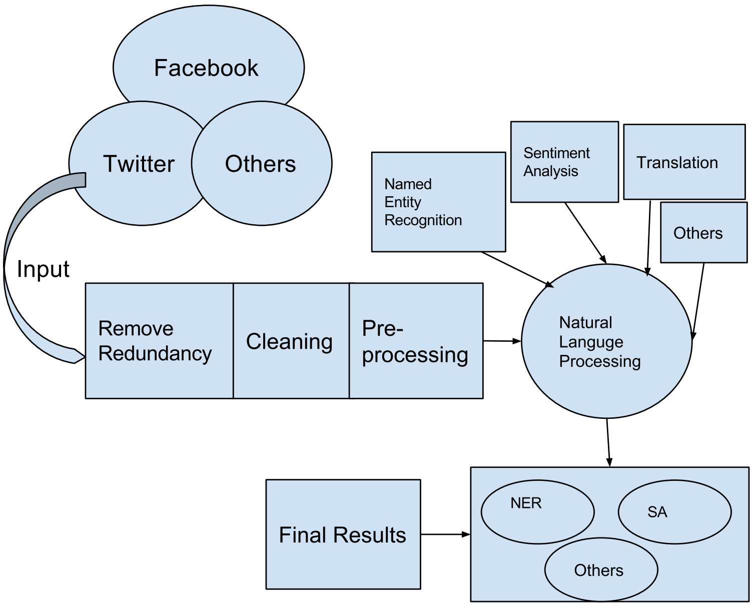


Figure 3.1: Automation of Decision Support System

which is taken as an input by the sentiment analyser. To reduce the performance degradation, it is necessary to pre-process data efficiently. Given the importance to minimize the human intervention in sentiment analysis and to get better results, systematized and efficient mechanisms are the need of the hour. Normalization is the basic task to handle performance degradation of various natural language processing tasks. The term 'Normalize' in past, is taken as to just make the content in a well structured format. These days normalize has broader term in the field of natural language processing. It includes handling slangs, spell correction, finding missing words, cleaning the text, etc. In this chapter, the system design and algorithm to handle unstructured or noisy data for sentiment analysis is presented. Although, the presented algorithm is generic in nature, it can be applied and tested it for sentiment analysis. A hybrid technique is proposed which comprises of two modules: corpus based and dictionary based. In corpus based modules, bigram and trigram word vector is used based on tf-idf [67]. It depends over the occurrence of these word combinations in the corpus. On the other hand, the data is normalized by eliminating the slangs, emoticons, noisy text, etc using dictionary. The generality of the algorithm makes it beneficial in various language processing tasks such as summarization, named entity recognization, etc.

3.2 Background

Researchers are working in the field of natural language processing for increasing the automation of decision support system. As people have much freedom to write over the web, the need to normalize their content also increased. In past, people prefer writing the text in the formal manner. With the rise in the web content, people prefer writing in short form, slangs, mistaken words, etc. It is needed to normalize the web content.

3.2.1 Pre-processing

Agarwal et.al.[6] normalized the text using character-level statistical machine translation system and training through a manually annotated dataset. From their work, it has been proved that automated normalization of data is more efficient than manual normalization. Dealing with slangs, was still in question. Their results were further modified by Baccianella et.al.[10]. They worked over the normalization of the textual data using weighted finite state transducers by using the phonemes. Results have been shown that their system outperformed the state of art machine translation. Later, due to the growth in free choice of writing over the internet people started writing in short forms. These short forms were taken as misspelled words by various language analysers. Baldwin et.al.[12] used the variations in spellings written by people for analysis. For normalization, training through the tool over human annotated dataset was used. They were also focused for correcting spelling for specific word i.e. keyword. Further, Baron et.al.[15] has proved the effectiveness of normalized text over the performance of text to speech system. Their research also helped normalization in gaining more attention. Cambria et. al.[22] presented an approach for normalization using machine translation. Correction of mistaken punctuation along with filling up with the missing words was included in their work. Results of proposed approach outperformed but its performance was completely depends over the translators effectiveness. Chieu et. al.[28] said in their work that normalization was not the matter of just replacing the words, it actually depends on the target application. System was designed by them to handle domain dependent normalization. Clavel et.al.[29] worked for performing all the necessary steps to have the formal structured content. They used the corpus based technique for short message normalization. In their work, translation of the short messages to formal English language was handled. Dey et.al. [38] and Ebrahimi et.al.[42] both normalized the text independent to the discipline in which it was used. Our approach is based on this fact to have a normalized text for general purpose. Apart from the above mentioned researchers, there are many other researchers who are working in the same field and a lot many to come.

3.2.2 Sentiment Analysis

Sentiment analysis is at the cross roads of Automatic Decision Support Systems, aims at finding the opinion regarding any entity by the web users. This work is proliferated with the rise in social media content and availability of writing freely over the internet. After efficient pre-processing of the text, we can apply sentiment analysis over the given documents. In literature, two kinds of approaches are conferred i.e. Corpus based and Dictionary Based. Both of these approaches has their own pros and cons. Corpus based approaches are basically depends upon the term frequency[67] value for positive and negative terms appeared in any document. Khan et.al.[59] used fine gained corpus for not only detecting the sentiment but also the implicit aspect and the global entity about which the sentiment have been generated. Researchers used camera and mobile phone reviews for their work. This also enhanced the work of implicit entity detection in sentiment analysis. In their work, the importance of corpus based sentiment analysis has been shown. Dictionary based approach is very significant in case of domain dependent sentiment analysis. The only drawback is to large volume of dictionary items give better results. Ljubesi et.al.[66] used domain dependent semantic orientation for sentiment analysis. Lexicon based approach for sentiment analysis was used by them. The investigation of the feature importance and contextual information in deducing the sentiment has done by them. Out-performance of the results of their system with respect to the state of art sentiment analysers was shown in their work. Agarwal et.al.[6] shown the significance of PoS tagging for feature selection. Lopes et.al.[67] identified the circumstances for the growth of individual advancement for cross-disciplinary work. Dictionary based approach was included. In their work, they primarily embedded artificial intelligence in the form of neural networks for opinion mining. Different agent based approaches were also employed. Their work results were significant in some areas but not generic in nature. Later, people work in finding the relation of various entities in the field of sentiment analysis. This further increases the need of having efficient sentiment analysers. Tsai et.al. [93] used random walk and iterative regression for first building concept level lexicon. They used commonsense for the annotation of the lexicons. Later, Desheng et al.[99] found the correlation between stock price and the reviews of stock finance through sentiment analysis. Their work highlights the need of accessing the reviews efficiently for automation of decision support system. Pennell et.al.[81] developed a strategy to extract sentiment from textual as well as visual web data in a combined way. The results has been shown better as compared to state-of-art sentiment analysis in Chinese language as well as visual sentiment analysis individually. In the field of linguistic analysis, emoticons were considered important as other context by Chen et.al.[26]. Cambria et.al.[22] highlighted various issues in sentiment analysis. Ebrahimi et.al.[42] used sentiment analysis in presidential election to check the popularity of the candidate. From the recent studies, it has been analyzed people use social data for various natural language processing tasks such as named entity recognition, text summarization , sentiment analysis, etc. This arised the need of effective processing of the data. Data over the web is very noisy i.e. contains emoticons, slangs, misspelled words, etc. For effective results, this noisy data needs to be normalized. Researchers use either dictionary based technique or corpus based approach to deal with the noisy data. This chapter proposes a novel hybrid approach which uses lexicon and corpus based approach in a combined manner.

3.3 Proposed Methodology

The hybrid framework of the proposed system for sentiment analysis primarily consist pre-processing and sentiment score calculation as shown in Figure 3.2 . Pre-processing further includes tokenization, cleaning of data and normalization. Out of these, normalization affects the results to a great deal. It consist two stages for normalization:

- Handling emoticons or slangs using pre-defined list of positive and negative emoticons along with cross word dictionary.
- Handling slangs using maximum likelihood ratio.

For normalization, a corpus based module and a dictionary based module as a preprocessing of the textual data is included. It is composed of rich vocabulary of slangs in the form of normalized cross word dictionary and a corpus based term frequency vector for bigrams and trigrams. The maximum likelihood of the next

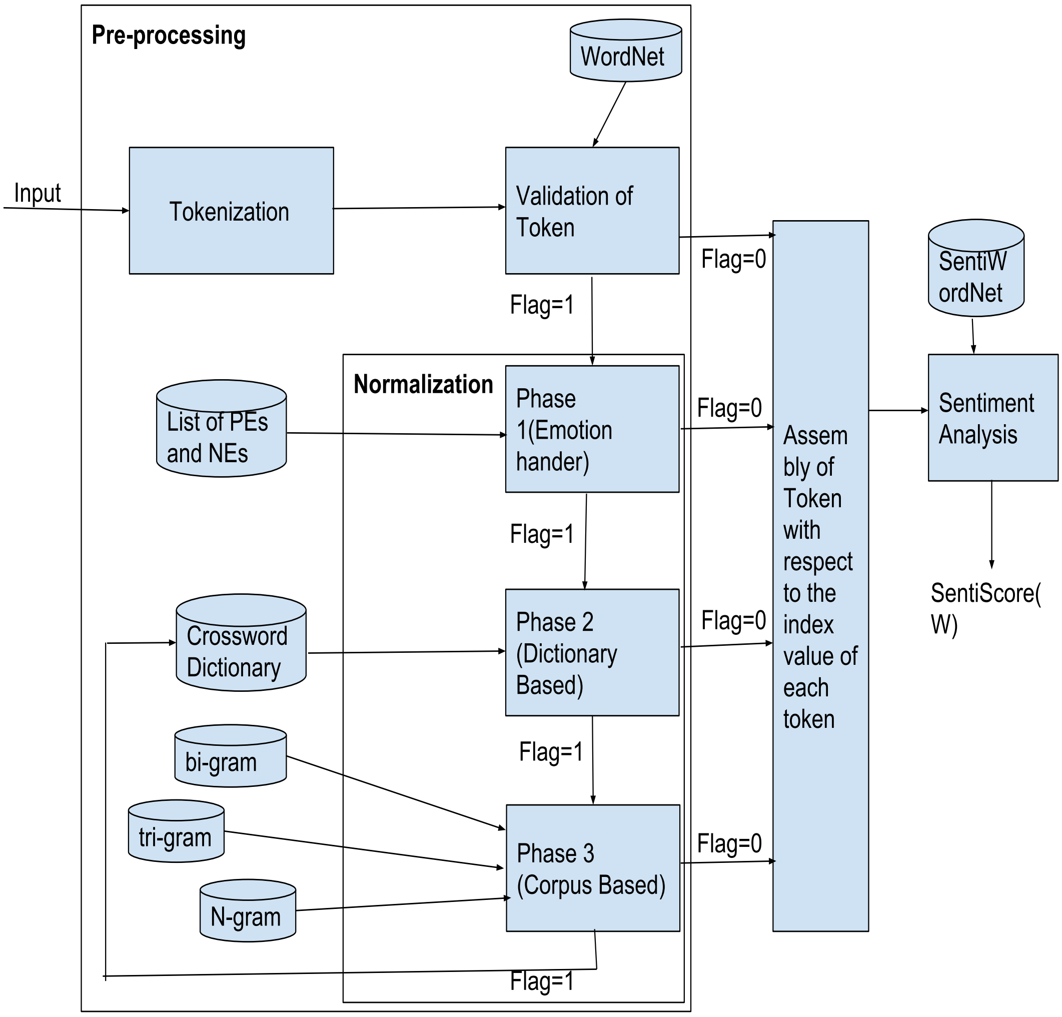


Figure 3.2: Proposed System for Preprocessing

word is calculated using maximum likelihood ratio. The proposed system is discussed in detail as follows:

3.3.1 Pre-processing of the Text

Pre-processing of the unstructured web contents is the major task to enhances the performance of sentiment analyser. It includes tokenization, normalization, etc.

Table 3.1: Tokenization

This is very gud phone and btry life is very long.
Tokenized text: This(1), is(2), very(3), gud(4), phone(5) , and(6), btry(7), life(8), is(9), very(10), long(11).
Input = 1 document.
Output = 11 tokens with their respective index values.

Tokenization

It is the process of breaking down the input into small units. The system divides the text based on space between words. It can be at word level, character level or sentence level. The proposed system uses word level tokenization. As shown in table 3.1

Validity

In this phase, the validity of the word is calculated using WordNet. Each token is passed for validity check. Here, tokens are searched from WordNet. If the token is found in WordNet, i.e. valid token(Flag =0) then it is directly passed to assembly phase. On the other hand, invalid tokens(Flag =1) are passed to normalization phase.

Normalization

Normalization of the web content basically includes 2 modules. These modules are explained in further sections individually as shown in table 3.2

Module 1: Invalid tokens are processed using dictionary based approach. The slangs, emoticons or noisy text is replaced using cross word dictionary. Here, positive and

Table 3.2: Segregation of Un-Normalized tokens

Input: Tokenized text : This(1), is(2), very(3), gud(4), phone(5) , and(6), btry(7), life(8), is(9), very(10), long(11).
Output 1(valid words as per wordnet): This(1), is(2), very(3), good(4), phone(5) , and(6), life(8), is(9), very(10), long(11).
Output 2(invalid words as per wordnet): btry(7)

negative emoticons are replaced with their respective meaning i.e. :) means happy, :(means sad and many more. If the replacement is found in the pre-defined dictionary, then the whole text is assembled in assembly phase. Phase 1 and Phase 2 from Figure 3.2 are included in module 1.

Module 2: In this, the corpus based approach is used to normalize the text. Bigrams and trigrams based on term frequency for the given corpus is used. Slangs in this module are corrected using point-wise mutual information (PMI). PMI [45] is calculated using equation 3.3.1. Phase 3 from Figure 2 describes the functionality of module 2.

If w1 is the word followed by incorrect word or the slang. w2 is the predecessor ofw1 in bigram trigram list.

$$PMI(\frac{t1}{t2}) = \log \frac{pr(t1 \wedge t2)}{Pr(t1)Pr(t2)} \tag{3.3.1}$$

Where,

Pr(t1 $\wedge$t2) is the actual co-occurrence probability of term1(t1) and term2(t2). Pr(t1)Pr(t2) is the co-occurrence probability of the two terms, if they are statistically independent. Maximum likelihood is calculated using PMI. Maximum likelihood [25] method is the

Table 3.3: Normalization Based on PMI

PMI(life—battery) = 0.32
PMI(life— animal) = 0.31
PMI(life—the) = 0.01
Maximum likelihood value = 0.32
Output : battery(7)

Table 3.4: Assembly of Tokens

Input (From module 1): This(1), is(2), very(3), good(4), phone(5) , and(6), life(8), is(9), very(10), long(11).
(From module 2): battery (7),
Output: This(1), is(2), very(3), good(4), phone(5) , and(6), battery(7), life(8), is(9), very(10), long(11)

procedure of finding the value of one or more parameters for a given statistic which makes the known the likelihood distribution a maximum. After calculating the maximum likelihood, the misspelled word is replaced with the word having maximum value of PMI. For this OvA (one-vs-all) strategy is used to handle the slangs as shown in table 3.3

3.3.2 Assembly of Tokens

In this phase, all valid tokens (flag=0) are ensembled with their respective index values. This is very important part of the system to retain the semantic orientation of the words. As if any negative pointer such as not is misplaced, it may give us wrong results. So, the words are studded as per their original index values. illustrated in table 3.4

3.3.3 Senti-Score Calculation

After the completion of normalization, the standard sentiment analysis algorithm is applied to the whole document. In this study, SentiWordnet [10] is used.

3.4 Proposed Algorithm

In this section, an algorithm for the proposed system is presented. For the proposed system, two algorithms are designed: Normalize text Algorithm(Proposed Algorithm 1), Senti-Strength/Sentiscore Algorithm (Proposed Algorithm 2) . Algorithm 1 is used to normalize the given documents. Algorithm 2 is used to find the exact classification of reviews in positive or negative category with the magnitude calculated using Senti-WordNet i.e. Sentiscore.

3.5 Experimental Results

3.5.1 Dataset

For the experimental setup, data is collected from the SMS Spam Collection v.1 corpus and blogs of 134 customers as used by Dey et.al.[38]. The corpus is a collection of 5,574 English messages. The corpus is representative sample for public data available over the web sphere. To build the dataset, blogs manually filtered by removing hashtags, hyperlinks, etc. For the analysis of results, Gold standard of the dataset is built by the linguistic experts.

Algorithm 1 Normalization()

```
Document(D)containing noisy data, where D = $d_1, d_2, d_3, ......., d_n$
'n' is the total number of documents
'T' is list of tokens, where T = $t_1, t_2, t_3, ....., t_m$
'm' is the total number of tokens in a document
'W' is the valid word found in WordNet
Initial list of Bi-grams and tri-grams
Initial list of Positive emoticons (PE) and negative emoticons(NE) with their corresponding meaning
Crossword Dictionary containing normalized slags ($C_d$)
List of assembled Words($L_w$)
$i \leftarrow 1$
$Tokenization$
for d$\in (d_1, d_2, d_3, ....., d_n)$ do
    Tokenize(T)
    for $t_i \in (t_1, t_2, t_3, ....., t_m)$ do
        if $t_i \in$ Standard Stop Words then
            Discard
        else if $t_i \in W$ then
            append $t_i$ to $L_w$
            Appnd $t_i$ with respective index value to $(L_w)$
        else if $t_i \in (PE \sqcup NE)$ then
            replace $t_i$ with normalized word from PE or NE
            Appnd $t_i$ with respective index value to $(L_w)$
        else if $t_i \in C_d$ then
            Apply maximum likelihood ratio based on pointwise
            mutual gain of bigram and trigram for slag replacement
            Appnd $t_i$ with respective index value to $(L_w)$
        end if
    end for
end for
Apply SentiScore($L_w$)
```

Algorithm 2 SentiScore()

```
INPUT:
Document(D)containing noisy data,
where D = $d_1, d_2, d_3, ......., d_n$
'n' is the total number of documents
OUTPUT:
$W_s$(Weighted SentiScore of each document/SentiScore)
{
Token list(T) = $t_1, t_2, t_3, ....., t_m$
'W' is the valid word found in WordNet
'q' is the total number of words in a document
'm' is the total number of tokens in a document
$'L'_n$ is the list of Positive words in a document
$'L'_p$ is the list of Negative words in a document
$'P'_w$ is the weight of positive term as per SentiWordNet
$'N'_w$ is the weight of negative term as per SentiWordNet
}
$j \leftarrow 1$
for $d_i \in D$ do
    Stemming
    Normalization
    Tokenize(T)
    for $t_i \in (t_1, t_2, t_3, ....., t_n)$ do
        if $(t_j \in W) \cap (t_j \in L_p)$ then
            $W_{pos}(j) = P_w(t_j)$
        else if $(t_j \in W) \cap (t_j \in L_n)$ then
            $W_{neg}(j) = N_w(t_j)$
        else if $t_i \in W \cap (t_j \in L_n) \cap (t_j \in L_p)$ then
            $W_{neu}(j) \leftarrow 0$
        end if
    end for
        $$W_s = \sum_{j=1}^{m} W_{pos}(j) \pm \sum_{j=1}^{m} W_{neg}(ij) \qquad (3.4.2)$$
end for
```

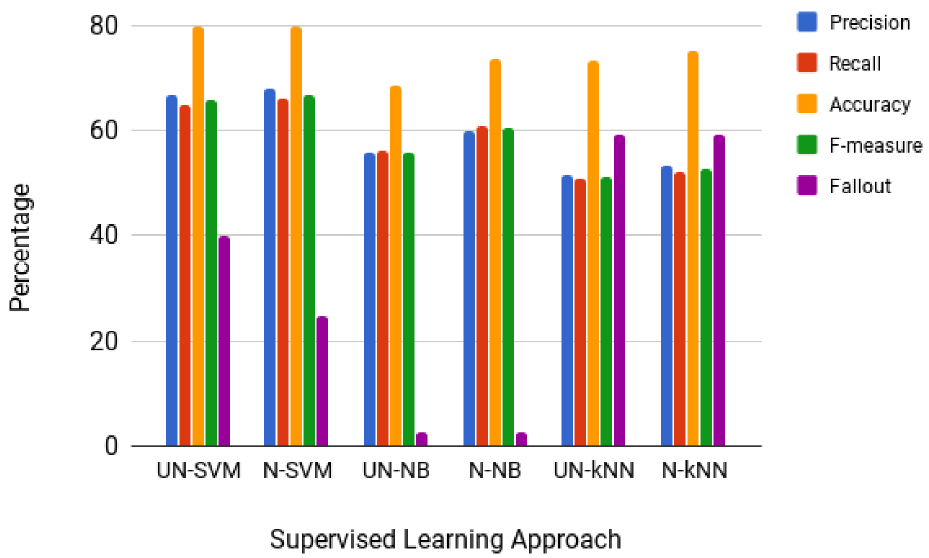


Figure 3.3: Comparison of Performance Based on Un-normalized(UN) and Normalized(N)Data

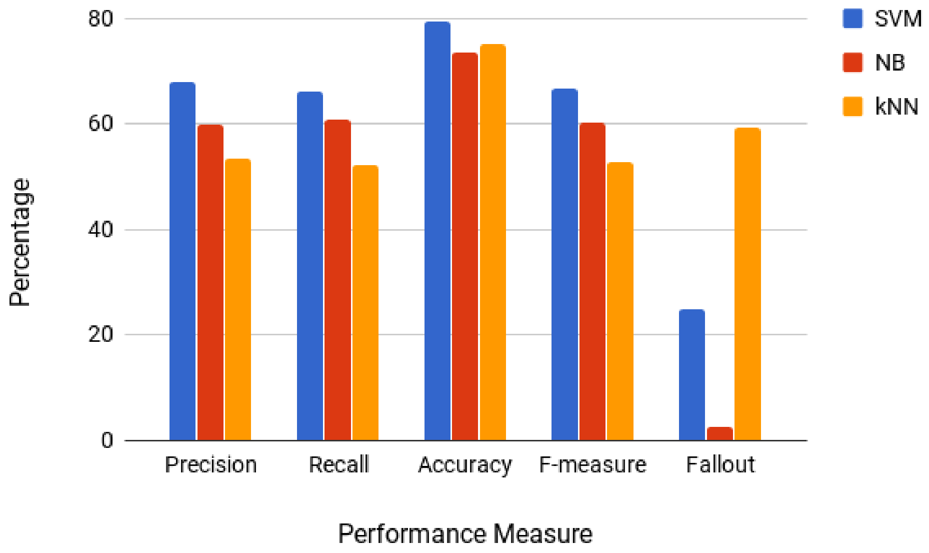
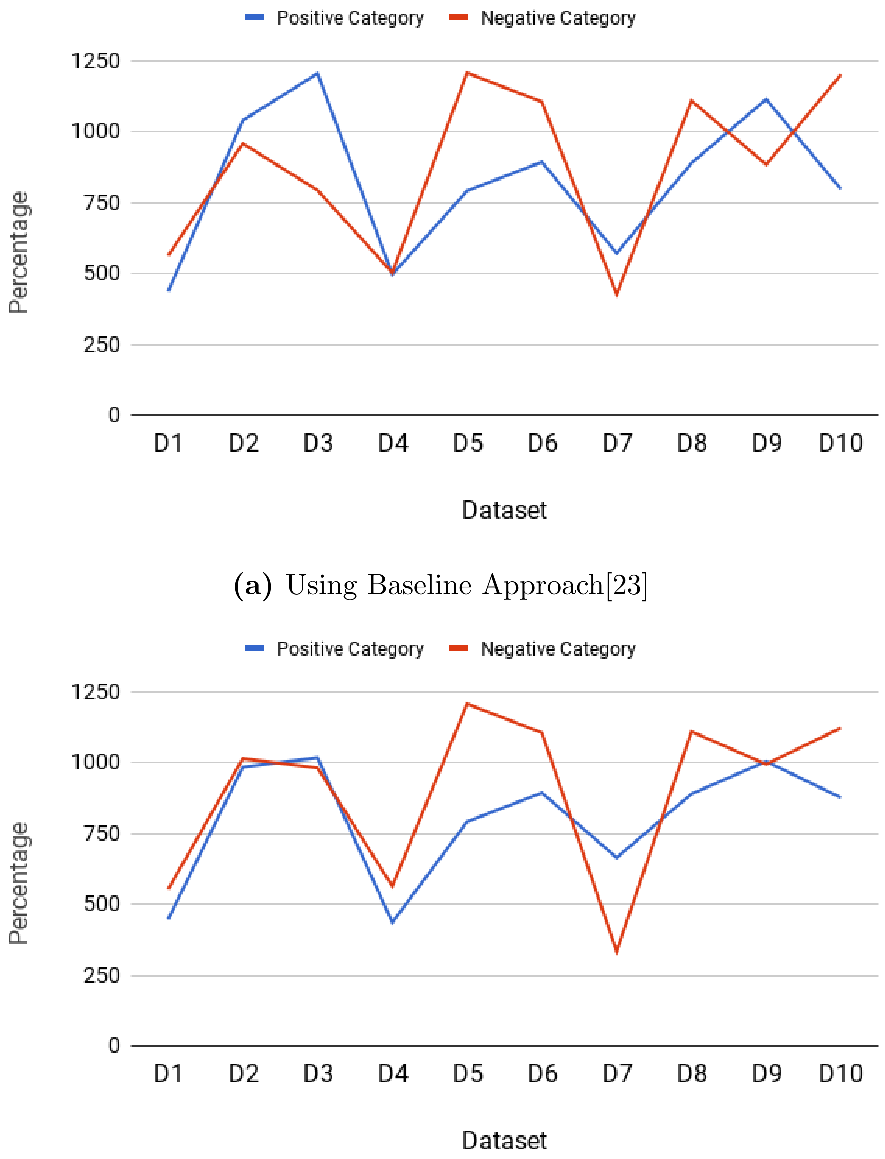


Figure 3.4: Performance Evaluation Based on Supervised Learning

3.5.2 Evaluation

In this section, exploration of results is done based on the automatic normalization of the content using hybrid approach. To examine the results, a common evaluation method as described in performance matrix of chapter 2 i.e. Recall, Precision, Accuracy, Fallout and F-measure are used. All the documents used for experimentation contained short messages which removes the need for applying dimension reduction approach to it. In table 3.5, the results are presented based on supervised approach, like SVM, Naive Bayes and k-NN. These results are divided into two categories: Normalized dataset and Un-normalized dataset. The results obtained by the proposed method are graphically shown in Figure 3.3. On comparison, we found better results for the normalized dataset than the un-normalized dataset. This shows the importance of normalization for any natural language processing task. Another significance of normalized data is reduction in fallout in case of SVM. For Naive Bayes and kNN , less significant results are observed in fallout for normalized and un-normalized datasets. On the contrary, this study has found the recall and precision values are raised in normalized data processing. Figure 3.4 shows the performance of various supervised approaches. It is found that results are better with SVM. It is noticeable that there is a trade-off between recall and fallout, precision and fallout. Naive Bayes has shown minimum fallout. The highest value of fallout in k-NN reduces its performance in some cases wherever less fallout value is appreciable. It depends on the application and domain to choose the learning algorithm.



(a) Using Baseline Approach[23]

(b) Using Proposed Approach

Figure 3.5: Classification of Reviews in Positive and Negative Category

Table 3.5: Results Based on Normalized and Un-normalized Datasets

Dataset	Model	Precision	Recall	Accuracy	F-measure	Fallout
Un-normalized	SVM	66.74	64.91	79.55	65.81	40
Un-normalized	NB	55.69	56.14	68.57	55.91	2.68
Un-normalized	k-NN	51.38	50.91	73.33	51.14	59.18
Normalized	SVM	67.95	66.14	79.55	66.79	24.77
Normalized	NB	59.93	60.68	73.44	60.33	2.45
Normalized	k-NN	53.29	52.05	75.16	52.66	59.02

Table 3.6: Accuracy of the Proposed Hybrid System

Approach	Accuracy
Dictionary based	85.01
Corpus based	84.03
Gold standard	88.13
Hybrid approach (proposed)	87.70

From table 3.6, it can be seen that the results are comparable with the rule based, corpus based and also with the Gold standard. After successful normalization using hybrid approach, sentiment analysis is applied over the given dataset. The variations in the result is shown in Figure 3.5. The dataset used is divided into small sets($D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_10$). Each set contains equal number of documents for removing biasing.

Figures clearly show that the normalization affects the results of sentiment analyzer to a great deal. After applying hybrid normalization on the datasets for sentiment

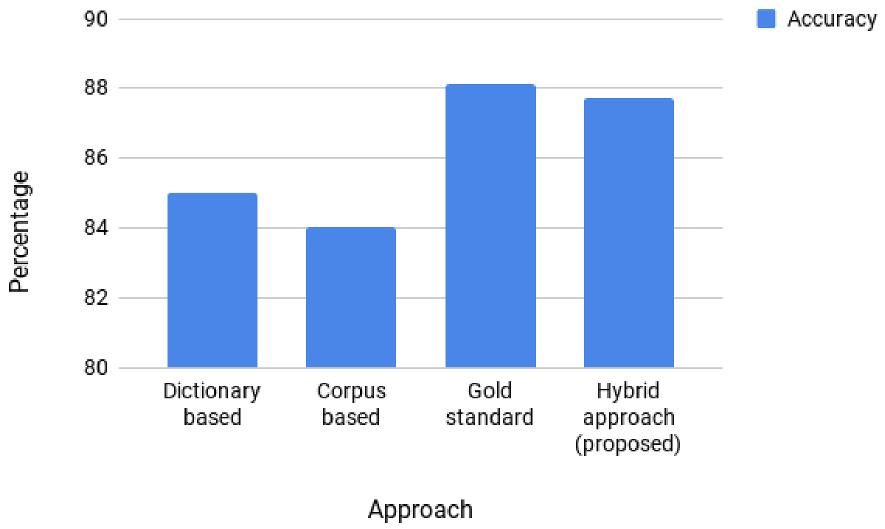


Figure 3.6: Effectiveness of Proposed Approach

analysis, it is found more documents fall in positive category shown in Figure 3.5b. These results may differ for different datasets. Results obtained after normalization are more realistic as it is more close to the Gold standard shown in Figure 3.6. It is found that hybrid approach based normalization gives better results than rule based and corpus based in terms of average accuracy.

3.6 Summary

In this chapter, a novel hybrid approach for normalization based on corpus as well as dictionary for various NLP tasks has been disussed. Experimental results have shown better performance for normalized data in terms for Recall, Precision, Accuracy and Fallout. The accuracy is more as compared to state-of-the-art dictionary or corpus based approaches. Although, our approach could not meet the Gold standard completely, it outperformed the existing techniques. Our proposed approach is generic in nature. Currently, it is applied for sentiment analysis to categorize the documents in positive or negative. It can be applied in other domains also.

Chapter 4

Normalization of Macaronic Content for Sentiment Analysis

In the former chapter, we have discussed the open rules set by online review communities that permit users to write opinions, queries and suggestions in unstructured/informal language over different online platforms. Each individual review plays a critical role in helping people make decision of buying or selling any service or product. These reviews are the primary feed of Automatic Decision Support Systems(ADSS) that perform on the basis of sentiment analysis. Now, the time comes which shows the need to elaborate the research in the field of multilinguality. People write more in their native language along with the base language of any document. In this chapter, the concern is about the macaronic content which is a form of multilingualism.

4.1 Introduction

Setting a uniform language for online review submission is a complicated task as the users are from different backgrounds. A report on online review processing concludes that over two third users on the internet are from a non-English background. The primary reason quoted behind the same is the proficiency of people to learn only two languages at a time, with perfection. However, the socioeconomic branch of online community support users to write in any language they know and prefer. A person belonging to different geographic areas have right to use their domestic language to write reviews online. This makes online review a multi-lingual text, a text comprising more than one type of languages. Similarly, any sentence having more than one language is known as macaronic text [84] i.e. Hinglish, Dunglish, etc.

Example 1: Airtel has अच्छा network.

The above mentioned sentence is a macaronic sentence, with both Hindi and English language words in it.

The linguistic variation of the online review data, makes the processing more complex. The lack of language resources over the internet also make it more complicated to deal with the diversity, altogether. On the other hand, it is important to consider reviewing from different users in different languages. This is the reason that there is a high demand for multilingual automated system. Derkacz et.al.[37] found necessary conditions for automation of multilinguality over the network. With advanced language processors, a multilingual system can be built, wherein, multilingual systems read the data of a multi-lingual text, whereas for macaronis text, the system reads the data of the complete line, word by word. In this chapter, the proposal for sentiment organizer is put forward that allows successful processing of macaronic text. Prior to processing reviews are to put into base language.

4.2 Related Work

Many researchers have contributed towards the field of natural language processing. Kaur et.al. [57] worked on the sentiment analysis of the reviews written and submitted in Punjabi language. The researcher compiled and segregated the reviews into positive and negative category. The researchers also highlighted the need of requiring lexicon. Das et.al.[33] concluded the research on the review written in Bengali Language. Das et.al.[32] analysed the need of SentiWordNet for Bengali language, this helped other researchers to perform more efficient sentiment analysis. Their task was based on the supervised learning, i.e. Support Vector Machine (SVM) with Bengali SentiWordNet. Along with this, researchers also represented feature extraction for Bengali language. Das et.al.[31] developed the subjectivity clues based on theme detection techniques. He used Bengali Corpus and later compared the outcomes with English Subjectivity detection. The researcher[34] also derived a gaming theory that helps researchers in building the SentiWordNet in any required language. However, for this, the assistance of the particular linguistic expert is a must. Joshi et.al.[56] applied supervised learning approach while working on a project using Hindi- SentiWordNet. To ensure that the polarity of the each document is preserved during translation, the researchers have used standard translation techniques.

Bakliwal et.al.[11] researched on subjectivity detection based on graph theory. An in-depth study allowed researchers to find out the effect of synonyms and antonyms over the subjective nature of the document. The results were applicable for both-Hindi and

English languages. Based on the study, the researchers claimed that the same theory can be used in other languages too. In other research, Das et.al.[35] established a system that help in deducing the emotion and their intensity through the hidden sentiments in any data through supervised learning methods. Richa et.al.[87] worked on a survey for sentiment analysis in Hindi language, wherein the results displayed that Hindi language analysis is more complex than English language analysis. The study indicates that the non-uniform nature of the Hindi language is the primary reason for this complexity. Researchers[87] established a system that indicates the polarity of various Hindi movie reviews. Parul et al.[9] composed a sentiment analyzer for studying the reviews written in Punjabi language based on different machine learning algorithms. Raksha et.al.[86] worked on semi-supervised technique for detecting the polarity in Hindi movie reviews. The researchers reported that the proposed system is 87% accurate when used on the basis of bootstrapping and graph-based approach for sentiment analysis. For determining the opinion orientation of reviews, Pooja et.al.[77] used Hindi SentiWordNet. All these findings were concluded using unsupervised learning. Kerstin et.al.[36] established a system that helps obtaining the polarity of reviews written in any language other than English, which is considered as one of the most resource-rich language till date. The two primary tools used by the researchers included standard translation methodology and supervised learning for sentiment analysis. C. Banea et.al.[14] worked and established a system that was based on translation of any input language other than English, through supervised learning approach. They used Google, Moses, Bing translators and more translators to accurately translate the text.

Table 4.1, presents a summary of the work done by various researchers. The table indicates the focus of researchers in the area of multilingual sentiment analysis. The primary focus was to translate the entire document after base language detection as an alternative to finding the language of the individual word. Often, this can lead to the elimination of an opinion bearing word in any foreign language.

Example 1, अच्छा’(in Hindi) which means ’good’(in English), might be eliminated in case the document language is detected as English. To ensure that the decision support system is efficient and just, such documents should be processed more efficiently.

Upon the initial research, researchers experienced the need of using SentiWordNet for almost all languages on global parameters. The task is a complicated one. The motivation used for the proposed system states that the existing system for multilingual sentiment analysis does not process macaronic data with efficiency. As the internet is experiencing a heavy rush of macaronic content, having an established system becomes imperative. There are several reasons there is huge macaronic content available over the internet, which includes:

1. Limited resources: To pursue sentiment analysis, it is important to have access to lexicons or data in a particular language. Each language model represents significant variations. This concludes that a single system cannot be used for

Table 4.1: State of the Art Multilingual Sentiment Analysis

Author	Work	Level	Language	Results	Technique	Corpus	Year
Danet et.al.[30]	Classification of reviews into positive or negative opinion	Document level	Punjabi	Accuracy = 75%	Machine Learning	Blogs	2014
Derkacz et.al.[36]	Classification of reviews into positive, negative , neutral or emotion (sad,happy,etc)	Document level	Bengali	Precision = 70.04% , Recall = 63.02%	Machine Learning	Custom Lexicon	2010
Das et.al.[34]	Document are separated based on Domain independent subjectivity and factual content	Sentence Level	Bengali	Precision = 70.04% , Recall = 63.02%	Machine Learning	Custom Lexicon	2009
Das et.al.[32]	Sentiment analysis of Hindi reviews,English reviews using Hindi SentiWordNet	Document Level	Hindi, English	Precision = 70.04% , Recall = 63.02%	Supervised	Movie reviews	2012
Joshi et.al.[56]	Subjectivity clues based on antonym and synonym using graph theory	Document Level	Hindi,English	Accuracy = 79%	Supervised	Movie reviews	2012
Sharma et.al.[86]	Polarity detection of movie reviews using unsupervised techniques	Sentence Level	Punjabi	NA	Unsupervised	Movie reviews	2015
Arora et.al.[9]	Sentiment orientation of reviews written in Hindi language	Document Level	Hindi	Precision = 70.04% , Recall = 63.02%	Unsupervised	Movie reviews	2014
Sharma et.al.[87]	Sentiment analysis using Semi-Supervised techniques	Document Level	Hindi	Accuracy = 87%	Semi-Supervised	Movie reviews	2014
Pandey et.al.[77]	Opinion orientation of Hindi movie reviews is deduced using Hindi-WordNet	Document Level	Hindi	NA	Unsupervised	Movie reviews	2015

all languages.

For instance, there is no provision of spaces in Chinese language model, whereas all other language systems use space as a technique of tokenization.

2. Absence of uniformity of languages: Different languages have different conventional structures. There cannot be one generic structure model to process different languages in a similar manner.

   For instance, the English language structure is based on Subject-Verb-Object (SVO) whereas, in the Hindi language model follows Subject-Object-Verb (SOV).

3. Freedom of use of Native language: Online applications have busted the geographical boundaries and there are multilingual followers of a single account over the web. At times, the account owners prefer writing reviews in their native languages to connect with their native followers. When an automated system is used for the pre-processing, it can eliminate these native words considering them as a foreign language. This can cause significant loss of meaningful words during the pre-processing phase.

   Example 2: सैमसंग has an affordable price.

   Here सैमसंग(in Hindi) represents Samsung(in English) and can be easily neglected by an English language based model for being a foreign language word. It will make it difficult to extract Samsung as an entity.

4. In order to seek attraction: Many times, people use multilingual content or

Table 4.2: Tokenization at Different Levels

Level of Processing	Number of Tokens
Sentence Level	2
Word Level	10
Character Level	53

fancy words for advertisements, names of institutions or other establishments to seek attention. This becomes a tough task for the processors, as it makes the web content complex. To eliminate any such confusion, it is important to develop an efficient system that can process the macaronic language content.

Example 3: सamसung (Samsung) has an affordable price.

सamsung (Samsung) has an affordable price.

मike (Mike) likes to play football.

Hence, from the above examples, Samsung, Mike is hard to detect as it is being neglected by chosen language model.

4.3 System Design

In the proposed system, the focus is put to normalize the multilingual content, specifically macaronic text(consists Hindi and English) and classify the reviews into binary category i.e., Positive and Negative as represented in Figure 4.1.The said system comprises of three major components as mentioned below:

1. Language Processing

Table 4.3: PoS Tagging Using NLTK and Stanford Tagger

Test Sentence	Pos tagging by NLTK tagger	Stanford tagger
मीड़िया गयान का एक अच्छा सरोत हैं	मीड़िया—NN गयान—:का—:एक—: अच्छा—सरोत—हैं—	मीड़िया/VBZ गयान/NNP का /NNP एक /NNP अच्छा/NNP सरोत /NNP हैं /NNP
media is अच्छा source of knowledge	media—NNS is—VBZ अच्छा—: source—NN of—IN knowledge—NN	media/NNS is/VBZ अच्छा/JJ source/NN of/IN knowledge/NN
मीड़िया गयान का एक good सरोत हैं	मीड़िया—NN गयान—:का—:एक—:good —JJ सरोत —हैं—	मीड़िया/VBZ गयान/NNP का /NNP एक /NNP good/JJ सरोत /NNP हैं /NNP
media गयान का एक अच्छा सरोत हैं	media—NNS गयान—:का—:एक—: अच्छा—सरोत—हैं—	media/NNS गयान/NNP का /NNP एक /NNP अच्छा/NNP सरोत /NNP हैं /NNP

2. Text Processing
3. Sentiment Analysis

1. Language processing is the primary component of the proposed system. The component carries out the process of tokenization, detecting the language and converting the tokens to the base language. The sub-components of the component are described below:

   (a) Tokenization: This is the primary step of all language processing tasks. In this process, a sequence of words, sentences or characters are fed as input

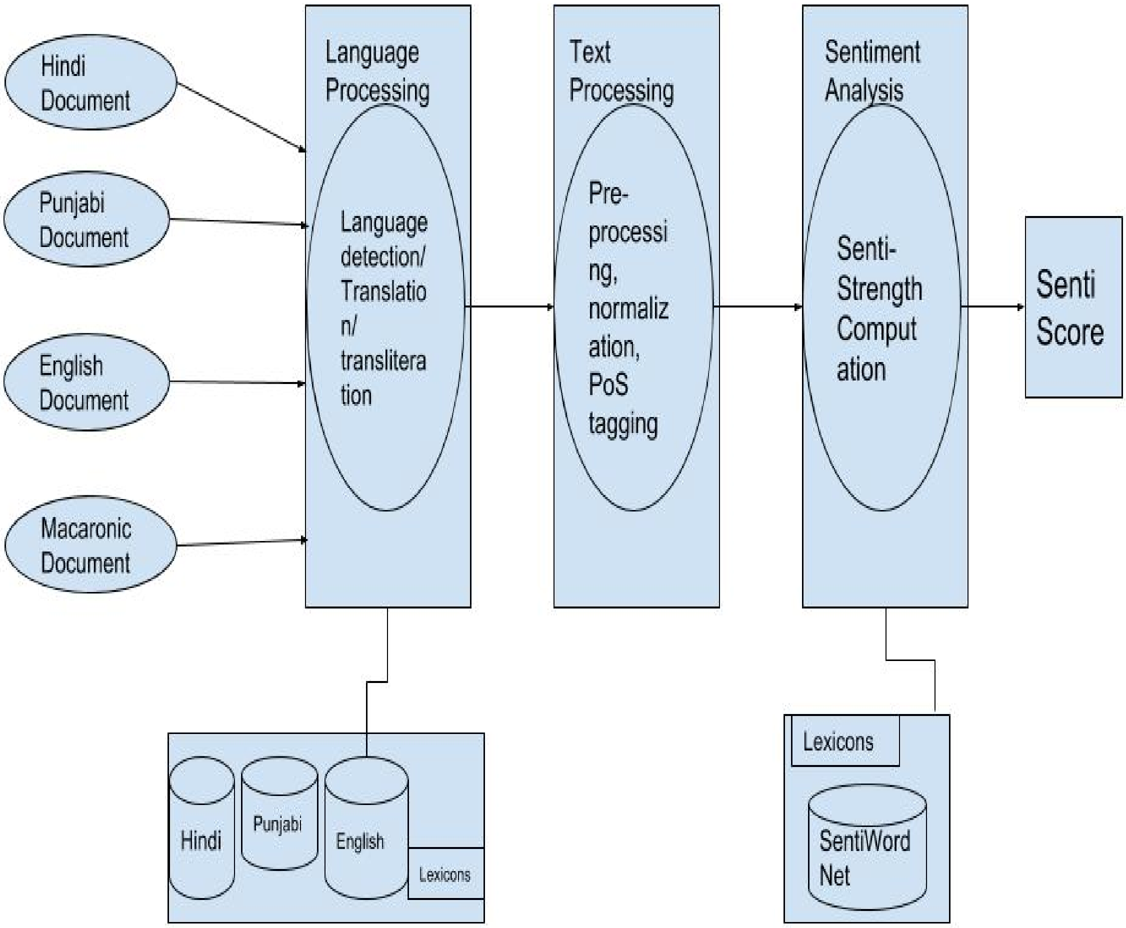


Figure 4.1: Proposed System Design

to a specific system. As a result, the process produces tokens. The use of the process at any level, i.e. sentence level, word level, character level, depends upon the granularity of the data. Number of tokens extracted for example 4 is represented in table 4.2. In the proposed system, the process of tokenization is used at the word level for macaronic language.

Example 4: People like Sony music player. It is of good Quality.

   (b) Language Detection/ Translation: PoS [47] tagging is the method used for language detection, represented in table 4.3. Various unrecognized or untagged tokens were passed through language detection module. The process produced tokens in the base language of the system. English is taken as a base language for the current study. In case any word is detected in Hindi WordNet, the translator would covert it from Hindi to English. Similarly, for Punjabi language, the word is translated from Punjabi to English. As a generic process used for all types of languages.

2. Text Processing: Text processing can be counted as the second imperative component of the proposed system. The various sub-tasks carried out at this level includes:

   (a) Normalization: Normalization is done once the filtration of subjective sentences is performed. In the process of normalization, all the grammatical variants of the sentences are regularized or processed. Past verbs (regular and irregular) / present verbs, classification of noun phrases in singular and plural are the popular grammatical variants. For efficient processing, data needs to be in a regularized format and this is the primary goal of normalization process. The process includes:

      i. Slangs handling: Slangs are an inseparable part of todays world and they play a significant role in mining the opinions. Rejecting all the slangs can be harmful to the study. Thus, researchers used various types of mechanisms in order to handle different types of slangs [17] as listed below:

- Emoticons: ':(', Bad, ':) happy .
- Interjections:Mmmmm-pleasure,hmmmm-wondering,etc
- Intensionally misspelled: cooooooool,goooooooood, nyt, etc
- Alphanumeric strings: gr8, 9t, etc.

Example 5: He is on cloud nine when his father gifted a new car. :)

He is on cloud nine when his father gifted a new car. Happy

Here, the emoticon :) is replaced with its meaning, i.e. Happy.

ii. Idiomization/Replacement of idioms with their actual meaning: A process for replacing idioms with the words of their actual meaning is Idiomization. Idioms are a critical part of the English language that helps in building opinions from a sentence regarding an entity. In case all the stop words are removed, it may affect the important part of the idiom. From example 5, the idiom (on cloud nine) is replaced with its meaning overjoyed, i.e, He is overjoyed when his father gifted a new car.

(b) Tokenization: The table 4.2 represents the output of the tokenizer at different levels . The primary reason behind this is that each word has to be processed as per its basic language.

(c) PoS Tagging: Part of Speech tagging is a critical aspect of natural Language processing tasks. At initial stage, the state of the art PoS taggers were used to check, if they were able to detect the foreign words or not. NLTK tagger [60] and Stanford Tagger[75] are the primary PoS taggers used for this study. The table 4.3 represents the test results of both the taggers on various sentences. There were several untagged tokens detected, which were then processed through language processing phase.

3. Sentiment Analysis: This phase analyse the potency of different reviews. It is measured in terms of sentiscore. The magnitude is calculated on the basis of sentiment associated with different documents which consist the reviews. For this purpose, SentiWordNet v3.0.0 is used. The Sentiscore of various reviews is listed in the table 4.4.

The above mentioned components are the basic building blocks of any sentiment analyser. Two basic algorithms are used for the processing of these components. Algorithm 3 is used to carry out the task of language detection. The language detection primarily focuses on normalizing the macaronic content to its base language. For the rest of two components, i.e. Text Processing and Sentiment Anslysis, algorithm 1 and 2 (discussed in chapter 3) is used, which is focused on normalizing the content to extract the SentiScore of all given documents. This is important to process sentiment analysis for documents in multilingual or macaronic language.

Algorithm 3 Macaronic Content handler()

```
Input: $Document\ D\ where\ D = d_1, d_2, d_3, ....., d_k$
        'k' is the total no. of documents
        'm' is the total number of words in a document
$L_s\ =\ language\ of\ segment$
$L_b\ =\ Base\ language\ (English)$
Output: $W_s(weightedSentiScoreofeachdocuemnt)$
Begin
for $k = 1$ to $k$ do
    $Tokenization$
    for $i = 1$ to $m$ do
        Encoding based on $UTF8$
    end for
    {Similar category segments are combined}
    Segmentation based on encoding.
    Language detection for each segment.
    if $L_s = L_b$ then
        $goto\ S1$
    else
        $Apply\ translation$
    end if
    S1  Assemble segments
    Compute SentiScore
end for
```

4.4 Evalaution

4.4.1 Dataset

The study is based on a corpus containing reviews of 10 movies 200 movie reviews i.e.100 positive and 100 negative. Out of these 200 reviews 160 were used for training, whereas 40 reviews were used for testing purpose. The reviews were long, ranging from 500 to 1000 words. The initial process of corpus classification was complicated as each review has to be rated and then classified into positive and negative. Reviews rated between 3 and 5 stars were are marked as positive and 0 and 2 are marked as negative. This classification is based on the assumption that the rating of the review relates to the sentiment associated with the review. The study has reviewed in multiple languages. These reviews contain words from more than one language, Hindi and English. Manual classification the reviews based on the type of language token. The primary rule for the classification was to retain the semantic structure of tokens. The gold standard was formulated by five graduate students performing the review process. Using the Kappa measure[21], we have performed the inter-personnel disagreement and the obtained a score of 0.61.

4.4.2 Performance

Formally, the efficiency of proposed sentiment analyser(PS) is composed of four tuples as described below:

PS{L,L_D,T,E_S}

Where, $'L'$ is a Learning Algorithm,

$'L'_D$ is Language Detection,

T is a Tagger,

$'E'_S$ is a Experimental Setup,

Here, the choice of optimal parameters corresponding to the factors mentioned above affects the performance of the analyzer. Sentiment analyzer offers maximum performance (PS_{max}), when used for optimal parameters choice. For the training purposes, machine translated data is used. For testing, learning algorithm based on the human classified dataset i.e. Gold Standard is used. The results obtained through a Sentiment Analyzer (PS) shows a negative effect because of the error in language detection phase E_{LD}, the same is shown in the equation 4.4.1 .

Table 4.4: Sentiscore Associated with Review

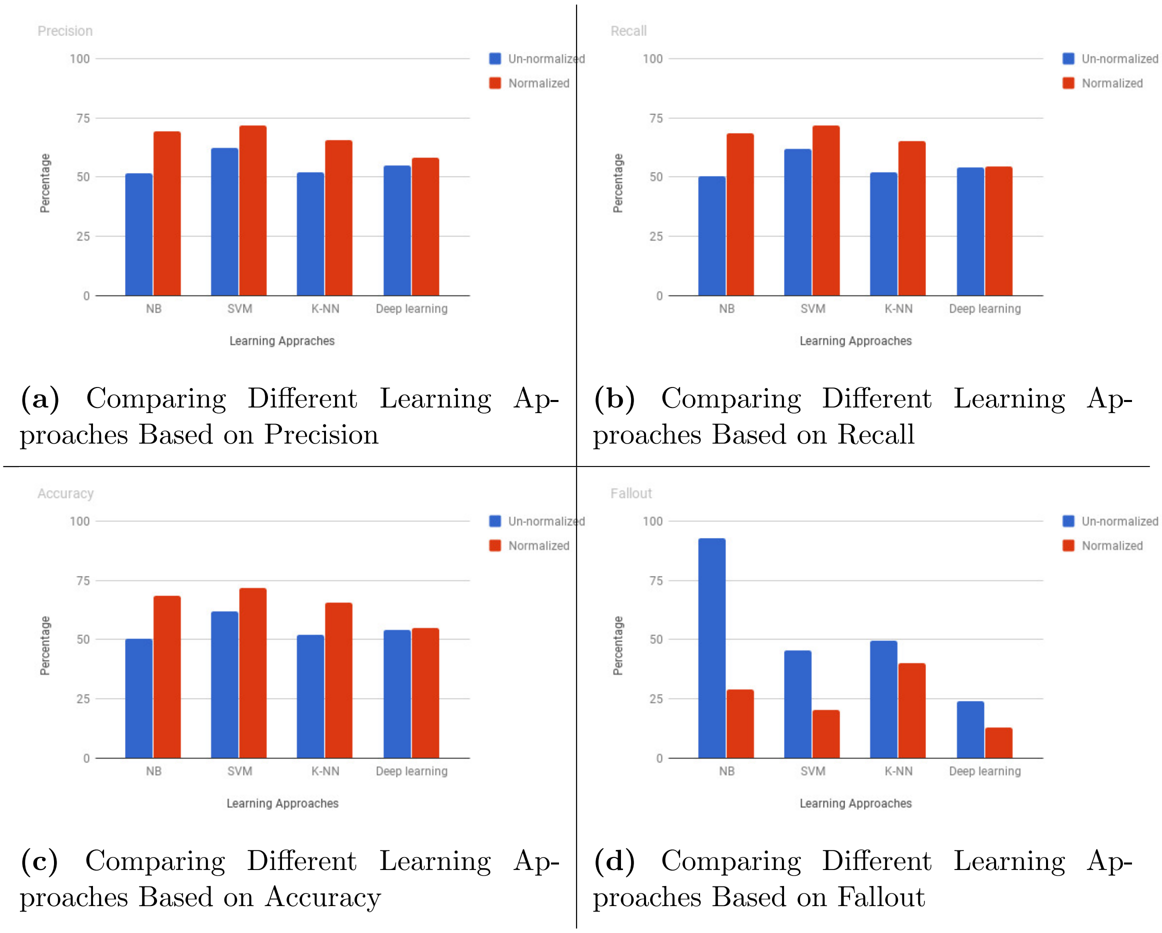
Test Sentence	SentiScore
texttt मीड़िया is good source of knowledge	0.47
media is good source of knowledge	0.47
मीड़िया गयान का एक अच्छा सरोत हैं	0
media is अच्छा source of knowledge	0
मीड़िया गयान का एक good सरोत हैं	0.47
media गयान का एक अच्छा सरोत हैं	0

Table 4.5: Un-normalized Macaronic Sentiment Analysis

Learning Approaches	Precision	Recall	Accuracy	Fallout	Time(sec)
NB	51.58	50.4	50.4	92.8	422
SVM	62.29	62	62	45.6	428
kNN	52.01	52	52	49.6	421
Convolutional network	54.96	54	54	24	751

$$PS = PS_{max} - E_{LD} \tag{4.4.1}$$

In case of optimal parameters, $E_{LD} \rightarrow 0$, PS = PS_{max}



(a) Comparing Different Learning Approaches Based on Precision

(b) Comparing Different Learning Approaches Based on Recall

(c) Comparing Different Learning Approaches Based on Accuracy

(d) Comparing Different Learning Approaches Based on Fallout

Figure 4.2: Comparison of Various Methods

Table 4.6: Proposed Normalized Macaronic Sentiment Analysis

Learning Approaches	Precision	Recall	Accuracy	Fallout	Time(sec)
NB	69.46	68.62	68.63	28.79	18
SVM	71.72	71.69	71.75	20.21	21
kNN	65.41	65.31	65.47	40.21	29
Convolutional network	58.03	54.56	55.00	13.04	440

4.4.3 Results and Analysis

Table 4.5 and Table 4.6 represents the results of our experimental study. The data represent that each machine learning approach represents different pros and cons. Precision, Recall, Accuracy, Fallout and Execution time are different aspects that are used for the evaluation. We have used 10-fold cross validation for validating the results. Support Vector Machines (SVM), Naive Bayes (NB), kNN and Convolutional network (Deep Learning) are a few tools that we have used during the experimental setup which helped us in analyzing the performance of the proposed algorithm. Table 4.5 and table 4.6 show the results of the process. Precision, Recall, Accuracy, Fallout is recorded in percentage and the time is recorded in seconds. Each learning technique takes different time for processing which depends upon data size, data types, number of columns, computer hardware, memory, background running processes, cores, etc. Table 4.5 and table4.6 has different entries which have helped in revealing the time trend corresponding to different learning models. Column named as t́imeíndicates that the reduction in the time comes to marginal levels in case of normalized content.

Table 4.7: Comparison with Existing Sentiment Analysis

Approach	Precision	Recall	Accuracy	Fallout
Baseline	55.21	54.6	54.6	53
Proposed	66.15	65.04	65.21	25.56

Based on time taken by various learning models, the order of performance from table 4.5 is given as:

$kNN < NaiveBayes < SVM < Convolutionalnetwork$

Based on time taken by various learning models, the order of performance from table 4.6 is given as:

$NaiveBayes < SVM < kNN < Convolutionalnetwork$

The figure 4.2 shows the result and indicates the performance of the proposed system through various learning approaches. The figure shows the performance of the proposed system based on different aspects. The proposed system functions well than state of the art analysers. Using Naive Bayes, it shows the rise in Precision and Recall by approximately 17.88% , 18.22% respectively. The results by other classifiers i.e. SVM, kNN and convolutional network, also show significant enhancement in the performance level. The results also indicate that there is a tradeoff between different aspects. For example, convolutional network shows more accuracy, but takes more time in comparison to other classifiers. The observation of figure4.2 also indicates a huge fluctuation in time taken by each classifier. It was observed that when proposed system is used, the training time is significantly reduced in each learning approach.

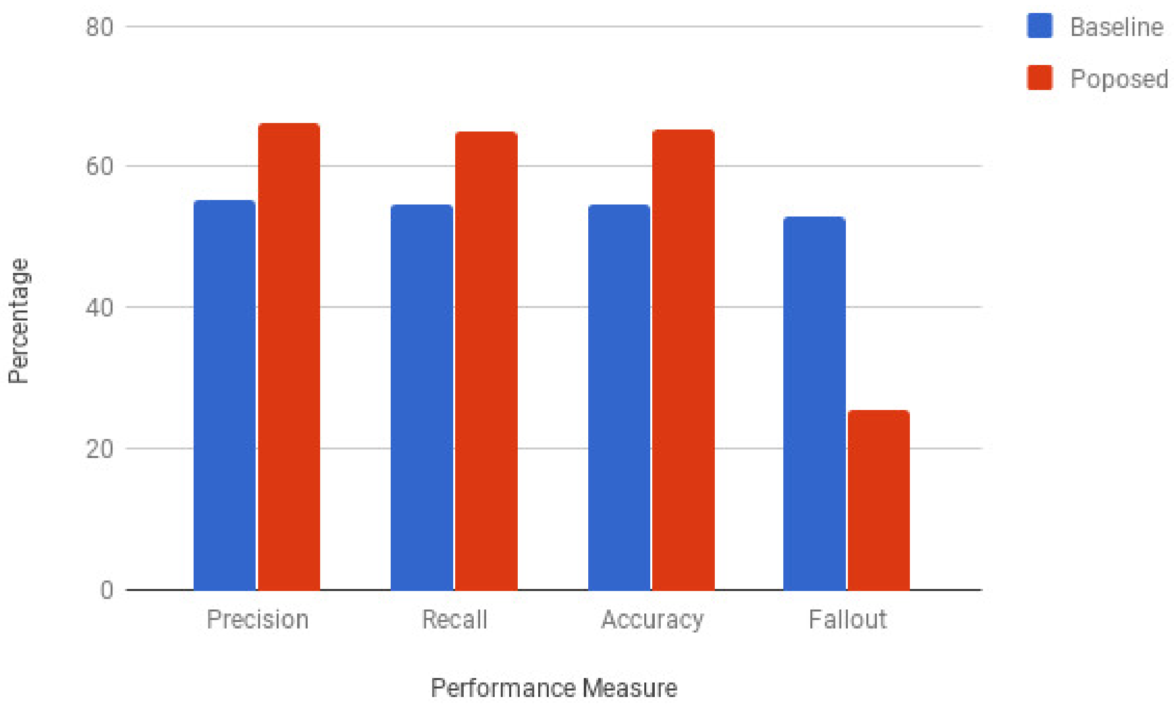


Figure 4.3: Effectiveness of the System w.r.t. Baseline Analysers

As we compared the data with other approaches, it was observed that the average value of precision, recall is increased and the fallout is decreased significantly. In figure 4.3 shows the effectiveness of the proposed system in comparison with the state of the art sentiment analysis for macaronic language.

4.5 Summary

There is a rising need for sensible computation of decision support system over the web, where a huge pile of user-generated content is already available. As more and more multilingual online content is submitted, the amount of web debris is rising, which is a primary factor affecting the results of decision support systems. In order to evaluate the negative trends and also propose a successful solution, this study focused on development of sentiment analysis based on the pile of words for macaronic reviews. Various supervised machine learning approaches were used and gave different cross-validated results. There was also the induction of training and testing from the field of machine learning. After prolonged analysis, we have concluded that the performance measures do not show any trade-off. However, the need for normalizing the content was observed throughout the study. During the study, sentiment analysis for macaronic text having words of both Hindi and English languages. On an average, a rise of 11% was noticed in precision and recall values. The study also showed that using the proposed approach, the training time can also be reduced significantly. We have further plans to use the proposed system to develop a more enhanced system of evaluating macaronic data that have used two or more languages. We also have plans to use our system of proposed algorithm for entity extraction.

Chapter 5

Handling Temporality for Query Based Sentiment Analysis

As discussed in the preceding chapters, nowadays a large number of users express their views for the products over various online platforms. These reviews are useful for potential consumers and manufacturers. Sometimes outdated reviews may result in biased sentiment analysis, which may or may not represent the current scenario. To remove this limitation, this study tries to implement temporal sentiment analysis of reviews by providing appropriate weightage to the reviews. In this chapter, the proposed algorithm addresses the challenge of real time query based sentiment analysis. The focus of this research is to devise an algorithm which gives the real essence of sentiments in the textual communication.

5.1 Introduction

Huge raw data available over the web, makes it a hard task to take any decision for any product automatically. Therefore, the automated analysis of product reviews based on natural language processing is of great value. This type of analysis is called as sentiment analysis or opinion mining. Sentiment analysis[79] (SA) is a quintuplet consisting, $'e'$ is an entity, $'a'$ is an aspect of the entity, $'s'$ is the sentiment on aspect, $'h'$ is opinion holder, $'t'$ is time of opinion as described in chapter 1.

i.e. SA = {e,a,s,h,t}

It involves building a system for collecting and categorizing the documents into positive, negative or neutral. Although, many researchers have worked to deal with various aspects of sentiment analysis. Still, many of the aspects need attention. Temporality is one of them. Time plays an inevitable role in all spheres of our lives, still real time has long been a forgotten dimension in state of the art sentiment analysers that perform automatic analysis of the reviews.Generally, with time the opinion of people is changed about any entity. Therefore, sentiment analysers should capture the temporal factor in the analysis. Present day sentiment analysers takes the time explicitly, i.e. date of post takes into consideration[76]. Temporality in real time sentiment analysis is achieved by formulating rules based on metadata as well as the linguistic context of words. Present sentiment analysers evaluate the overall Sentiscore[91] of any entity irrespective to the document creation time or review posted time. SentiStrength[79] uses a lexical approach that exploits a list of sentiment-related terms and standard linguistic rules and methods to express sentiment. It generates Sentiscore by giving equal weightage to all the reviews without considering the temporal aspect. Therefore, takes the outdated reviews equally important as the present day reviews for the Sentiscore generation process. This somehow degrades the reliability of the sentiment analysers because the importance of bygone reviews varies with time depending on the query.

For Example, Query 1: Is ford car a good car?

For above mentioned query, the reviews of present (2017/2018) should be given more weightage. The reviews from the years other than 2017/2018 should be taken as bygone or outdated reviews. Hence, less importance should be given to these.

Query 2: Was ford car a good car in 2016?

In this, the reviews of present (2017/2018) should be considered unimportant. These may be assigned zero weightage in the overall opinion generation. On the other hand, the reviews of 2016 should be given high weightage. It was observed that there is a need of analysers which can work with temporality in sentiment generation. In this study, the proposed technique generates Sentiscore by focusing on temporality. It uses a linguistic approach to exploit the temporal behaviour of words along with metadata. This type of Sentiscore generated is termed as Tempo-Sentiscore.

5.2 Related Study

Many researchers worked in the area of sentiment analysis or opinion mining. Thelwall et.al.[92] developed various algorithms for the identification of subjective or objective nature of the textual data. These subjective clues are further classified as positive or negative. They developed algorithm to detect sentiment score in addition to sentiment polarity from the structured sentences. Strapparava et.al. [89] worked for the differentiation of emotions as mild or strong, this is same as deduced by humans.

With the passage of time, researchers found much of the textual data is informal and unstructured in nature. The need to tackle with this informal text for sentiment analysis was aroused. Thelwall et.al.[92] devised a technique for calculating the actual sentiment score from unstructured and informal text, i.e. short sentences. Most of the sentiment analysis work is based on WordNet[69] which is the electronic dictionary used for various linguistic tasks and SentiWordNet[10] is another lexicon which holds a numeric value given to various words contributing for calculating the actual magnitude or strength of the opinion. Temporal properties in natural language processing has not gotten the proper attention in sentiment analysis. Inclusion of time in the field of sentiment analysis also made it more valuable in decision support systems. Temporal aspect in sentiment analysis based on metadata (explicit) was considered by O'Connor[76]. They counted all instances of positive-sentiment and negative-sentiment through topic keyword during the specific time as mentioned explicitly in the query. Thelwall et.al.[90] in their work considered time associated for the analysis of sentiments. They have found the popularity of any event by taking time, event and its corresponding sentiment from the online reviews. Again in their work, they used the metadata for each review i.e. date of post. Han et.al.[51] gave a two-level constraint-based framework, one is for processing and second is based on reasoning over temporal information in natural language. Chang et.al.[24] devised an algorithm for temporal tagging, which not only recognise, but also normalize temporal expressions in English. Tempo-WordNet[39] has generated using various linguistic rules for the classification of sentences as temporal/Atemporal[39]. Sentiment analysis based on the temporal nature of the document considering metadata along with linguistic rules is yet not considered by the existing systems. Razavi et.al.[83] worked for deducing the sentiment of the text containing the dreaming content. They used short textual data for sentiment analysis of dreams. Fukuhara et.al.[47] proposed a sentiment analyser for analyzing temporal trends of sentiments and topics from texts with respect to time. They had shown the impact of sentiment corresponding to a particular topic at a specific time. In their work, they had used various news articles and data collected from weblogs. Other researchers[106] used images which contained geographical information. They had used that information for embedding temporality and for sentiment analysis, they used images. The effect of temporality was analysed[106] by showing their effect over communities. Along with it, the effect of preprocessing was also associated with sentiment analysis[50]. Results were found better in case of SVM from the state of the art. To reduce high dimensionality in processing through bag-of-words, a system was proposed[5], which minimizes the dimensionality by eliminating irrelevant features and noisy text. To broaden the work in area of sentiment analysis, Arabic social media data[4] was considered. Again in their research they had taken care of temporality of the reviews. Recently, temporal characteristics[27] has used for sentiment analysis of travel blogs over time, i.e. explicit temporality. Temporal sentiment analysis has used for person recommendation[49]. The brief description of sentiment analysis based on explicit and implicit temporal tag is given in table 5.1. It can be summarized from table 5.1 that temporal expressions hidden in the linguistic context of words are not considered by sentiment analysers. Although researchers consider explicit time in the form of metadata to deduce the sentiments. We have proposed a system that generates the sentiment analysis of the documents based on metadata as well as temporality of the word linguistically .The architecture of the system is as shown in figure 5.1.

For Example:

The lens quality was good. Posted on: 26/06/2018 It contains two aspects:

- According to the metadata (date of post), the given review is in present.
- According to the linguistic rules (was,were, etc - past), here the presence of the word was make the given review fall in past category.

Table 5.1: Summarization of Explicit and Implicit Temporality

Sr. No.	**Author**	**Explicitly mention of topic keyword**	**Implicitly deduce topic keyword**	**Handling the geographical dispersion of time**	**Fore-cast analysis**	**Weightage hinged to reviews w.r.t. time**
1	O'Connor et.al. [76]	Yes	No	No	Yes	No
2	Thelwall et.al. [91]	Yes	No	Yes	No	No
3	Razavi et.al. [83]	Yes	No	No	No	No
4	Dias et.al. [39]	Yes	Yes	No	No	No
5	Fukuhara et.al. [47]	Yes	No	No	No	No

5.3 Proposed Definition of Temposentiscore

It is the measure of subjectivity and opinion from the textual data. It usually captures a modified potency of the Sentiscore. It is a triplet comprising, s is Sentiscore associated with each document, t is the temporal tag (present, past or future) based on implicit and explicit tag assigned to each document, c is the weightage given to each temporal tag, i.e. c ∈ c1,c2,c3 where c1 is weightage given to present, c2 is the weightage given to past and c3 is weightage given to the future.

i.e. TS= s,t,c

The aggregated Tempo-Sentiscore of any entity is defined by the equation 5.3.1.

$$\begin{aligned} TS &= \frac{\sum_{i=1}^{n} Sentiscore(Temp_{Tag})_{present} * c1}{n} + \frac{\sum_{i=1}^{n} Sentiscore(Temp_{Tag})_{past} * c2}{n} \\ &+ \frac{\sum_{i=1}^{n} Sentiscore(Temp_{Tag})_{future} * c3}{n} \end{aligned} \tag{5.3.1}$$

where,

'Sentiscore$(Temp_{tag})'_{present}$ magnitude of the sentiment score by the base algorithm for the document in present.

'Sentiscore$(Temp_{tag})'_{past}$ magnitude of the sentiment score by the base algorithm for the documents in past.

'Sentiscore$(Temp_{tag})'_{future}$ magnitude of the sentiment score by the base algorithm for the documents in future.

c1,c2 and c3 are the variables whose value depends on the temporal(present/past/future) tag of the document. From equation 5.3.1, Tempo$-Sentiscore(TS)$ is calculated based on the formulated rules for the temporal tag mentioned in Table 5.2 . It includes T_{tag} and D_{tag} based on linguistic rules and metadata respectively.

5.4 Proposed System Design

The methodology of the proposed system involves :

1. Tokenization

2. Tagging

3. SentiScore Generation

4. Tempo-SentiScore Generation

The detail description of each is as follows:

5.4.1 Tokenization

Pre-processing of the text includes tokenization. It is a process of dividing the whole text into segments based on word boundaries or a delimiter depending on the language used. It binds the characters into semantic units. The origin of this approach is from Penn Treebank Project[69]

Test sentence:

He is very happy with the products of this company.

Tokens Generated:

He

is

very

happy

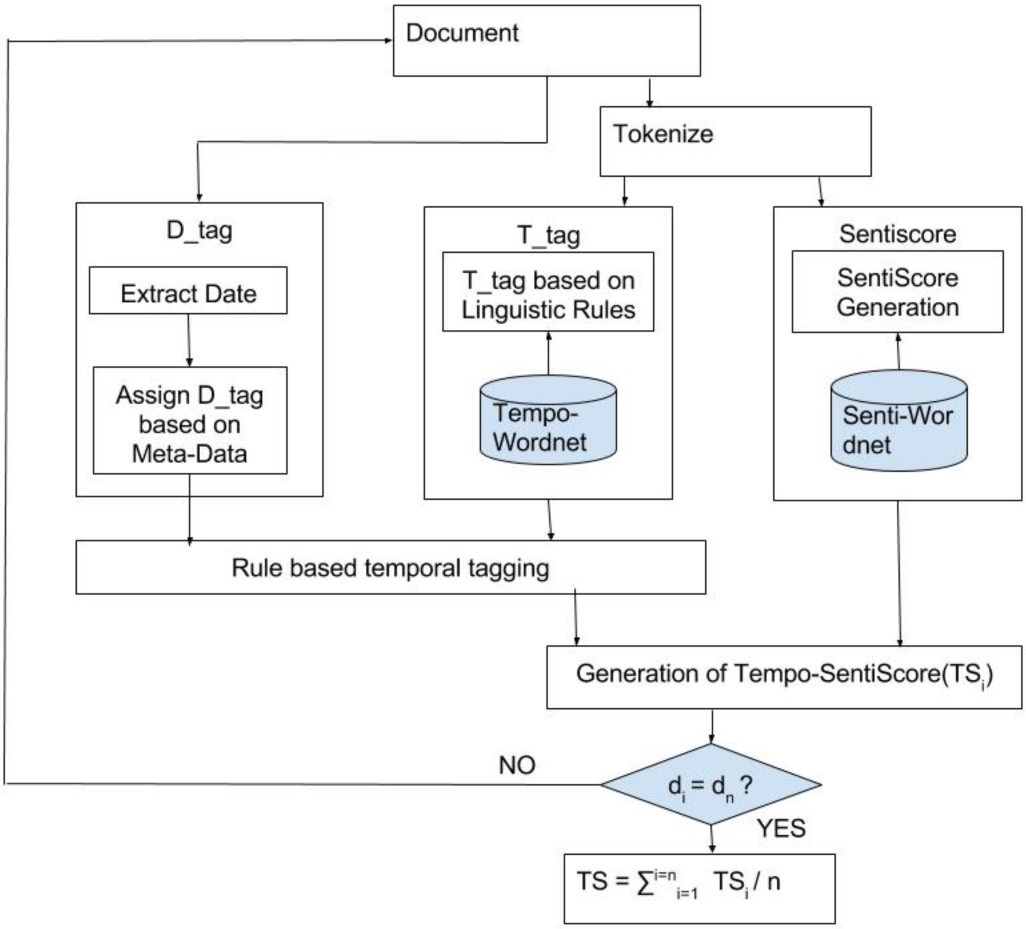

Document
Tokenize
D_tag
Extract Date
Assign D_tag based on Meta-Data
T_tag
T_tag based on Linguistic Rules
Tempo-Wordnet
Sentiscore
SentiScore Generation
Senti-Wor dnet
Rule based temporal tagging
Generation of Tempo-SentiScore(TS_i)
NO
$d_i = d_n$?
YES
$TS = \sum_{i=1}^{i=n} TS_i / n$

Figure 5.1: Proposed System Design

with

the

products

of

this

company.

5.4.2 Tagging

It comprises of 3 different phases,

- Implicit Tagging
- Explicit Tagging
- Generation of $Temp_{tag}$

Implicit Tagging T_{tag}: In this phase the temporal tag is assigned to documents based on various linguistic rules. The linguistic rules targeting the most used general terms of temporal expressions. TempoWordNet is used for having $Temp_{tag}$[11] i.e.

$was/were/had/...etcpast$

$is/am/are/...etcpresent$

$will/shall/...etcfuture$

For example:

Review 1 : The seats of this car was very comfortable. Posted on :02-05-2018

T_{tag} = past (implicit)

For above mentioned query, the review talks about the past opinion of the car. Although it is posted in present.

Review 2 : The camera of this mobile is awesome. Posted on:08-06-2016

T_{tag} = present (implicit)

Linguistically the post is considered as present opinion. However, it is written in the past year. Explicit Tagging D_{tag}: It is metadata based tagging phase. In this phase, the D_{tag} is assigned to each document. It is based on the document creation time or date of post. If the number of days exceeds the given threshold value, then D_{tag} is present and if it is below the threshold value then D_{tag} is past. The reviews from the year 2017 and 2018 are taken as present before 2017 every review is counted in the past.

For example:

The seats of this car is very comfortable. Posted on :02-05-2015

D_{tag} = past (explicit)

In the above example, this review is considered in past category.

The camera of this mobile is awesome. Posted on:08-06-2018

D_{tag} = present (explicit)

According to the date of post of the review, the given review is in present. Generation of $Temp_{tag}$: Rules for the $Temp_{tag}$ are formulated with the help of three linguistic experts. They were instructed to read D_{tag} and T_{tag} to annonate Temp_tag as present, Past and Future. With the rules as defined in table 5.2 , Temporal tag is assigned to each document. Rules for the assignment of temporal tag to each document Semantic Structure of rules:

ruleType: "tokens",pattern: ($/D_{tag}?/$,$/T_{tag}?/$,$Temp_{tag}$:?)

Where,

$'?'$ is replaced by present, past and future as by different temporal expressions.

Algorithm 4 Tempo-Sentiscore

```
Input: $DocumentD, whereD = d_1, d_2, d_3, ....., d_k$
        'n' is the total no. of documents
        'm' is the total no. of words
        'c1' is a weightage given to document of Present category, i.e. $Temp_{tag}$= Present and 0<= $c1$ <= 1
        'c2' is a weightage given to document of Past category, i.e. $Temp_{tag}$ = Past and 0<= $c2$ <= 1
        'c3' is a weightage given to document of Future category, i.e. $Temp_{tag}$ = Future and 0<= $c3$ <= 1
Output: TempoSentiSroce ($TS$)
Begin
for $i = 1$ to n do
    $Tokenization$
    for $k = 1$ to $m$ do
        Apply the linguistic rules for temporal tagging using TempoWordNet
        Calculate frequency count (Present, Past and Future)
        $T_{tag}$= max(Present, Past and Future)
        $D_{tag}$= Present or Past
        if T$_{tag}$ = NULL then
            $Temp_{tag} = D_{tag}$
        end if
        Calculate the Sentiscore 'W$_i$'
        if $T_{tag}$ = Present then
            if $D_{tag}$ = Past then
                $Temp_{tag}$ = Past
            else if $D_{tag}$ = Present then
                $Temp_{tag}$ = Present
            end if
        end if
        if $T_{tag}$ = Past then
            if $D_{tag}$ = Past then
                $Temp_{tag}$ = Past
            else if $D_{tag}$ = Present then
                $Temp_{tag}$ = Past
            end if
        end if
        if $T_{tag}$ = Future then
            if $D_{tag}$ = Past then
                $Temp_{tag}$ = Past
            else if $D_{tag}$ = Present then
                $Temp_{tag}$ = Future
            end if
        end if
        if $Temp_{tag}$ = Present then
            t$_i$= c1*w$_i$
        else if $Temp_{tag}$ = Past then
            t$_i$= c2*w$_i$
        else if $Temp_{tag}$ = Future then
            t$_i$= c3*w$_i$
            i = i+1
        end if
        if $i <= n$ then
            goto step 1
        end if
    end for
end for
```

$$Tempo - Sentiscore(TS) = \sum_{i=1}^{n} \frac{t_i}{n} \tag{5.4.2}$$

Table 5.2: Rules for Temporal Tag

Rules	T_{tag}	D_{tag}	$Temp_{tag}$
Rule 1	Present	Past	Past
Rule 2	Present	Present	Present
Rule 3	Past	Past	Past
Rule4	Past	Present	Past
Rule 5	Future	Past	Past
Rule 6	Future	Present	Future

5.4.3 Sentiscore Generation

SentiWordNet is the base for getting the actual magnitude of the sentiment for a document. For our work, we have used SentiWordNet[10]. The Sentiscore(w_i) is calculated using the Senti-score algorithm [91], i.e. Algorithm 2(as discussed in chapter 3).

Test sentence: Her happiness is increased by having this cellphone.

Sentiscore(w) = 0.271, After the completion of all the three phases, we get the following:

1. Sentiscore.

2. D$_{tag}$ (metadata based)

3. T$_{tag}$ (linguistic rules)

4. Temp$_{tag}$

5.4.4 Tempo-Sentiscore Generation

In this step, the Tempo$-Sentiscore$ is to be calculated which captures the real essence of the sentiment. The Tempo$-Sentiscore$ is calculated from equation 5.4.3.

$$TS_i = w_i * c \tag{5.4.3}$$

where, TS_i is the new magnitude of the opinion according to the temporal tag for it̂h document.

w_i holds the magnitude of Sentiscore by using SentiWordNet for ith document.

'c' , value of c depends on the temporal tag, i.e. present/past/future

For Example:

If Temp_tag p̄resent then c = c1

If Temp_tag p̄ast, then c= c2

If Temp_tag f̄uture, then c= c3

c1+c2+c3 = 1

Finally, Tempo-Sentiscore of any entity is calculated as an aggregation of Tempo-Sentiscore associated with each document (i) using equation 5.4.4.

$$TS = \frac{\sum_{i=1}^{n} TS_i}{n} \tag{5.4.4}$$

5.5 Proposed Algorithm

Based on the above discussion, an algorithm is devised as algorithm 4. The input is a collection of documents and the output is a Tempo-Sentiscore (TS) for an entity.

The algorithm is simplified for presentation clarity.

5.6 Experimental Setup

The experiment is held at document level. It contains the following components to deploy the proposed algorithm.

5.6.1 Dataset

We have used the standard dataset[16] .The version of the dataset[48] consists 2745 reviews of Ford car collected during the year 2007, 2008 and 2009. These reviews are collected from the social sites, i.e. Edmunds. In our proposed algorithm, we assumed the reviews collected in 2009 as present for temporal tagging based on metadata. It contains the review, date of post and entity about which the review is expressed. As shown in table 5.3, column heading D_{tag} which is based on metadata or date of post., T_{tag} values are based on linguistic context of words, $temp_{tag}$ contains the total number of review fall in each category (present/past/future).

Table 5.3: Total Number of Reviews in Each Category

Category	D_{tag}	T_{tag}	$Temp_{tag}$
Past	1970	279	2074
Present	775	2459	669
Future	0	7	2
Total	2745	2745	2745

In column D_{tag}, we have found the 775 reviews of present and 1970 reviews of past category. We found no review in future category. The second column (T_{tag}) gave the number of reviews categorized as in present, past and future based on the linguistic context of the words using table 5.3. The number of reviews in present, past and future are 2459, 279 and 7 respectively. It shows that most of the people prefer writing the reviews in present. We figured out from table 5.3 that number of reviews under column head ($Temp_{tag}$) changed in each category. It is observed that the consideration of linguistic context of words along with the metadata decreased the number of reviews in present, i.e. 669 and future reviews, i.e.2.

On the other hand, decreased the number of reviews fall in past category i.e. 2074.The reason is that some of these reviews were written in 2007 or 2008 which makes them to be considered in past. So, in the last column, we found the number of reviews in present and future category is reduced.

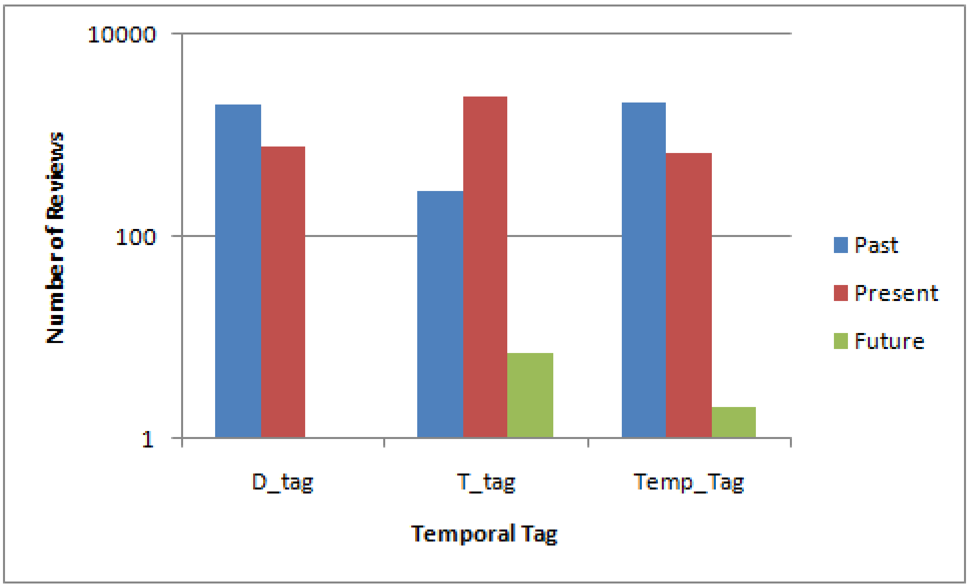


Figure 5.2: Total Number of Reviews in Present,Past and Future

The reviews fall in the present category of metadata (D_tag) are filtered by reducing those reviews which linguistically (T_{tag}) talked about the past status of ford car. The number of reviews is reduced to 669. Therefore, 106 reviews actually talking about the past status of the Ford car is shown in table 5.3. In figure 5.2, initially according to the date of post, the reviews in the present category were 775. By analyzing these reviews thoroughly based on the rules mentioned in table 5.2, we found that only 669 reviews are actually talked about the present scenario of the ford. For past the total number of reviews posted before 2009 were 1970. The number of reviews in this category was increased to 2074. We have found that 104 reviews which are posted in 2009 i.e. present actually described the past opinion towards the ford car. In future category, 7 reviews were found under T_{tag}. Further, applying the rules mention in table 5.2, only 2 reviews are actually considered as predictive opinion i.e. future. The next phase after categorization of reviews, is to generate the Tempo-Sentiscore. From table 5.3, it can be seen that the data in each category are different. To reduce the biasing between each class, we manually designed the dataset consists of 300 reviews in each category (present, past or future), i.e. 900 reviews. We took the help of3 linguistic experts to form the Gold standard.

Table 5.4: Results Based on Temporal Sentiment Analysis

Category	**R**ecall	**P**recision	**F**-measure
Present	92.66	92.33	92.49
Past	97.58	97.58	97.58
Future	82.35	83.05	82.69
Overall	90.86	90.98	90.92

5.6.2 Performance Evaluation

We used precision and recall to measure the performance of our method.

Table 5.5: Temporal Sentiment Analysis for Ford Car

Category	**R**ecall	**P**recision	**F**-measure
Ford07	Past	0.57	0.19
Ford07	Present	0.2	0.11
Real Time Sentiscore	Past + Present	0.37	0.08

The performance of the system for categorization of data in each class is based on the overlapping of the results with the Gold standard. The results are shown in form of Precision, Recall and F-measure in table 5.4 . Precision, recall and F- measure for individual class is described in table 5.4 . The average Precision and Recall of the proposed system over the given dataset is 90.98 and 90.86 respectively. The average F-measure is also calculated as 90.92.

5.6.3 Effectiveness of TempoSentiscore

Tempo-Sentiscore is calculated using equation 5.3.1 . Sentiscore is calculated by the sentiment classification through term scoring using SentiWordNet. The values of c1,c2 and c3 are gathered by the survey of more than 300 responses. The average of the weightage given by human annotators for c1,c2 and c3, i.e. present(c1), past(c2) or future(c3) is taken. The value of c1, c2 and c3 are found as 0.75, 0.15 and 0.10 respectively as per the average of their respective values collected by the survey. The values of c1, c2 and c3 are query based. For the query asking for the present status of the ford car, the weightage given to the present reviews is more than past or outdated reviews.

Query 1: Are people happy with the ford car?

All the reviews are taken for the analysis. Equal weightage is given to all the reviews irrespective of the temporal nature of the document.

Query 2: Were people happy with the ford car till 2007?

In this, the reviews till 2007 are gathered. The reviews of previous years (....., 2005, 2006) are taken as past, 2007 reviews are taken as present, while the reviews of 2007 pointing towards the future are taken as future.

5.6.4 Experimental Results

The experiment for the Tempo-Sentiscore task was carried out using the manual annotation by different linguistic experts. As for the Tempo-Sentiscore, there was not any kind of baseline for the comparison. This arises the difficulty in evaluating the performance. To evaluate the effectiveness of the proposed system, we employed the experiments to overlap with the Gold standard. The variation in the Sentiscore

Table 5.6: Assumption of Star Rating

Range of Sentiscore/TempoSentiscore	Rating
0-0.2	*
0.3 - 0.4	**
0.5 - 0.6	***
0.7 - 0.8	****
0.9 - 1	*****

i.e. Tempo-Sentiscore according to temporal tags is shown in table 5.5. For the star rating of any entity, Sentiscore plays a vital role. From table 5.6 , it can be seen that the star rating is affected by real time Sentiscore to a great extent. If the real time Sentiscore is used without temporal tagging, then ford07 is rated as **. On the other hand, with temporal tagging, it is rated as *.

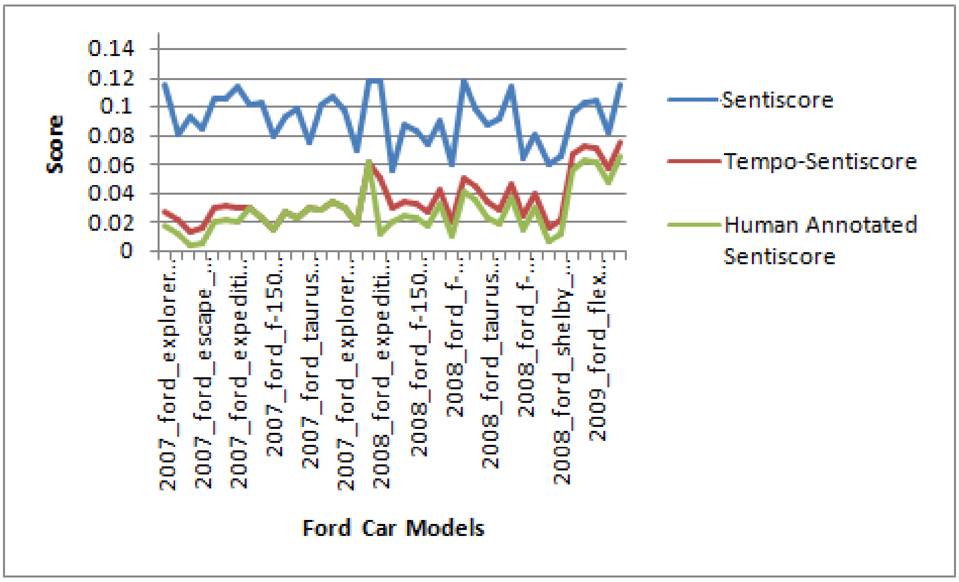


Figure 5.3: Tempo-Sentiscore Vs State of the Art Sentiment Analyser

The trend followed by both the Sentiscore and Tempo-Sentiscore is almost the same. The magnitude of the Tempo-Sentiscore is low as compared to the Sentiscore as shown in figure 5.3 . Linguistic experts agreement helped in showing the Tempo-Sentiscore represent the real scenario of the opinion about any entity. From figure 5.3, it is noticed that Tempo-Sentiscore is very near to the Human annotated results as compared to Sentiscore. It shows the reliability of the proposed algorithm.

5.7 Summary

In this chapter, we applied sentiment analysis methodologies to English. We have developed an algorithm for the generation of TempoSentiscore for efficient analysis of reviews to increase the reliability of the decision support system. Categorization of reviews for temporal tagging based on metadata and linguistic context of words in English language significantly outperformed. Our approach to evaluating Tempo-Sentiscore achieved high performance levels. Suggesting that these Tempo-Sentiscore methodology may be used in future by various sentiment analysers. The current performance of the system is promising for effective analysis of the reviews. In a nutshell, it is concluded that the Tempo-Sentiscore really affects the magnitude of overall opinion.

Chapter 6

Conclusions and Future Scope

The overall goal of our research work was to establish a pure connection between different aspects of sentiment generation. As the social media has a vast source of information, there should be a reliable senti-score generation methodology. The whole research work that has been carried out throughout this thesis is recapitulated in this chapter. This chapter firstly focuses on the summarise of the work. In later sections, it briefing the learning outcomes of the whole study, which further can be associated with state of the art research for the betterment of decision support system.

6.1 Thesis Summary

In chapter 1, the very nature of the sentiment analysis has discussed. The evolution of the sentiment analysis in the field of natural language processing is also exposed. The applications of this study are elaborated in this chapter. Chapter 2 presented the noble contribution of researchers in sentiment analysis. Chapter 3 addressed the issue of preprocessing for sentiment analysis of unstructured online reviews. It primarily focused on handling emoticons as well as slangs. It has shown the effectiveness of preprocessing in sentiment analysis. Chapter 4 unveiled the issue of processing macaronic content found over the web by sentiment analysers. For this, Hinglish (combination of

Hindi and English) is taken into account. Chapter 5 identified the need of capturing temporality in sentiment analysis. The combination implicit and explicit temporality gave rise to the formulation of new temporal tag. Based on this temporal tag, a new term is coined called as Temposentiscore. It also focused on how hidden temporality affects the star rating of a product.

6.2 Concluding Remarks

This chapter summarizes the resulting outcomes of the study presented in chapter 3, 4 and 5. The concluding remarks are fundamentally gained from three problems handling unstructured data, deal with Macaronic content and taking Implicit temporality into consideration for sentiment analysis, which is the articulation of presenting the study.

6.2.1 Pre-processing

People, these days have the flexibility to write and process different kinds of social data. Various decision support systems become the regulatory bodies for automatic processing of social data. This huge data consists unstructured content comprising slangs, emoticons or misspelled data. Various natural language processing (NLP) tasks are carried out to feed into computerized decision support systems. Among these, sentiment analysis is gaining more attention. These systems aim to aid decision making for customers, manufacturers, etc. by providing easily accessible information when needed. To enhance the performance of decision support system, it is very much required to process data efficiently. The proposed approach for normalization has shown better experimental results. Although, this approach could not meet the Gold standard completely, it outperformed the existing techniques. Currently, it is used for sentiment analysis to categorize the documents in positive or negative. Detail is discussed in chapter 3.

6.2.2 Macaronic Content Sentiment Analyser

Language is the main element of communication. Nowadays, the language diversity is very high over the internet. People use different languages for communication over the internet. This type of multilingual communication affects a large-scale and smaller businesses. So, to effectively process the web data for sentiment analysis, a need of handling macaronic content aroused. In this study, we have proposed an effective methodology to process the documents consists two languages. Hindi and English is considered for the study. It shows better results for sentiment analysis presented in chapter 4.

6.2.3 Temporal Sentiment Analysis

With such an amazing growth, today business more or less depends on social media or web. Target audience is hanging around the social media. People remain busy with social networks. Social media help people to sell or buy any product , use any services , etc through the review analysis. During such an analysis, time plays an inevitable role. If the analysis give an obsolete analysis. It hinders the performance of automatic decision support system. the analysis mainly focuses on the fact that which reviews or content needs to be considered. Along with it, the importance of present, past or future aspect of the content also needs to be considered. The affect of temporality is considered in review analysis in this study as discussed in Chapter 5.

6.3 Contribution

1. Increase in Reliability : We have determined that the contribution of objectives incense the reliability of the system to a great deal. It includes trimming to unwanted data or obsolete data upto an appropriate level. In some cases, less obsoletes data gets less importance as per the present data used for processing. The aim of presented work is to generate a sentiment for automated decision support system based on temporality. The reliability of the decision support system is increased as compared to the state of the art. The proposed temposentiscore is reliable.

2. Effective Preprocessing : Normalization has been used in data preprocessing before passing to sentiment analyser is also illustrated in this work. We have analysed how stop words filteration affect the results. Preprocessing also includes to deal with unstructured data contains slangs, syntactically weak structure.

3. Domain Independence : The objectives associated with this work is domain

independent. Algorithms are domain independent.Processing of macaronic content for sentiment generation is also taken into account. Classification of multilingual reviews has been studied in this regard and found the results are better in the proposed approach. Our methods are domain-independent and robust.

4. Tempo-Sentiscore : In this work, we have coined a term Tempo-sentiscore which captures the temporality hinged with each review. This tempo-sentiscore is better than sentiscore in terms of automated decision support systems.

6.4 Future Work

1. Extending Lexicons: The obvious way to improve a lexicon driven approach as presented in chapter 3 is ofcourse to utilize more social data. So, the proposed approach for normalization of noisy data, i.e. handling slangs, emoticons, etc needs to have extended versions.

2. Increase the level of granularity of temporal aspect: Linguistic temporality can be captured at higher levels of granularity. For the betterment of state of the art sentiment analysers temporality much be captured at other levels also.

3. Association of more than two languages: In this study, we have been discussing macaronic sentiment analysis by taking two languages into account. In futute, for the intense analysis more than two languages need to be covered.

4. Integration with other systems: We have applied normalization, macaronic content handling for sentiment analysis. In future, the proposed algorithms can be applied to other domains also like named entity recognition, summarization, etc.

5. Capturing the intensification: In literature, we have found researchers worked on long tail words. These words sometimes capture a intensity of emotion. We need to capture this intensity for sentiment analysis.

Bibliography

[1] http://www.internetlivestats.com/one-second/. Accessed July 4 , 2018.

[2] https://www.ibm.com/analytics/. Accessed June 17, 2018.

[3] https://risk.lexisnexis.com/our-technology/. Accessed April 22, 2018.

[4] Abdul-Mageed, M., Diab, M., and Kübler, S. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language 28*, 1 (2014), 20–37.

[5] Agarwal, B., and Mittal, N. Machine learning approach for sentiment analysis. In *Prominent feature extraction for sentiment analysis*. Springer, 2016, pp. 21–45.

[6] Agarwal, B., Mittal, N., Bansal, P., and Garg, S. Sentiment analysis using common-sense and context information. *Computational intelligence and neuroscience 2015* (2015), 30.

[7] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, pp. 487–499.

[8] An, N. T. T., and Hagiwara, M. Adjective-based estimation of short sentences impression. In *KEER2014. Proceedings of the 5th Kanesi Engineering*

and Emotion Research; International Conference; Linköping; Sweden; June 11-13 (2014), no. 100, Linköping University Electronic Press, pp. 1219–1234.

[9] Arora, P., and Kaur, B. Sentiment analysis of political reviews in punjabi language. *International Journal of Computer Applications 126*, 14 (2015).

[10] Baccianella, S., Esuli, A., and Sebastiani, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC* (2010), vol. 10, pp. 2200–2204.

[11] Bakliwal, A., Arora, P., and Varma, V. Hindi subjective lexicon: A lexical resource for hindi polarity classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)* (2012), pp. 1189–1196.

[12] Baldwin, T., and Li, Y. An in-depth analysis of the effect of text normalization in social media. In *HLT-NAACL* (2015), pp. 420–429.

[13] Banea, C., Mihalcea, R., and Wiebe, J. Multilingual sentiment and subjectivity analysis. *Multilingual natural language processing 6* (2011), 1–19.

[14] Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2008), Association for Computational Linguistics, pp. 127–135.

[15] Baron, A., and Rayson, P. Automatic standardisation of texts containing spelling variation: How much training data do you need? 1–25.

[16] Bashir, S., Afzal, W., and Baig, A. R. Opinion-based entity ranking using learning to rank. *Applied Soft Computing 38* (2016), 151–163.

[17] Boiy, E., and Moens, M.-F. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval 12*, 5 (2009), 526–558.

[18] Bollegala, D., Mu, T., and Goulermas, J. Y. Cross-domain sentiment classification using sentiment sensitive embeddings. *IEEE Transactions on Knowledge and Data Engineering 28*, 2 (2016), 398–410.

[19] Brill, E. Some advances in transformation-based part of speech tagging. *arXiv preprint cmp-lg/9406010* (1994).

[20] Brody, S., and Diakopoulos, N. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the conference on empirical methods in natural language processing* (2011), Association for Computational Linguistics, pp. 562–570.

[21] Bunt, H., Petukhova, V., Traum, D., and Alexandersson, J. Dialogue act annotation with the iso 24617-2 standard. In *Multimodal Interaction with W3C Standards*. Springer, 2017, pp. 109–135.

[22] Cambria, E., Poria, S., Gelbukh, A., and Thelwall, M. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems 32*, 6 (2017), 74–80.

[23] Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., and Haruechaiyasak, C. Discovering consumer insight from twitter via sentiment analysis. *J. UCS 18*, 8 (2012), 973–992.

[24] Chang, A. X., and Manning, C. D. Sutime: Evaluation in tempeval-3. In *SemEval@ NAACL-HLT* (2013), pp. 78–82.

[25] Charpentier, A. Maximum likelihood estimates for multivariate distributions. *Machine Learning 14* (2018), 05.

[26] Chen, C.-H., Lee, W.-P., and Hwang, J.-Y. Tracking and recognizing emotions in short text messages from online chatting services. *Information Processing & Management* (2018).

[27] Cheng, C., and Xu, J. A sentiment analysis model based on temporal characteristics of travel blogs. *Data Analysis and Knowledge Discovery 1*, 2 (2017), 87–95.

[28] Chieu, H. L., and Ng, H. T. A maximum entropy approach to information extraction from semi-structured and free text. *Aaai/iaai 2002* (2002), 786–791.

[29] Clavel, C., and Callejas, Z. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing 7*, 1 (2016), 74–93.

[30] Danet, B., and Herring, S. C. Introduction: The multilingual internet. *Journal of Computer-Mediated Communication 9*, 1 (2003), 0–0.

[31] Das, A., and Bandyopadhyay, S. Theme detection an exploration of opinion subjectivity. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (2009), IEEE, pp. 1–6.

[32] Das, A., and Bandyopadhyay, S. Opinion-polarity identification in bengali. In *International Conference on Computer Processing of Oriental Languages* (2010), pp. 169–182.

[33] Das, A., and Bandyopadhyay, S. Sentiwordnet for bangla. *Knowledge Sharing Event-4: Task 2* (2010).

[34] Das, A., and Bandyopadhyay, S. Sentiwordnet for indian languages. *Asian Federation for Natural Language Processing, China* (2010), 56–63.

[35] Das, D., and Bandyopadhyay, S. Labeling emotion in bengali blog corpus– a fine grained tagging at sentence level. In *Proceedings of the 8th Workshop on Asian Language Resources* (2010), p. 47.

[36] Denecke, K. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on* (2008), IEEE, pp. 507–512.

[37] Derkacz, J., Leszczuk, M., Grega, M., Koźbiał, A., and Smaïli, K. Definition of requirements for accessing multilingual information and opinions. In *Multimedia and Network Information Systems*. Springer, 2017, pp. 273–282.

[38] Dey, L., and Haque, S. M. Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition (IJDAR) 12*, 3 (2009), 205–226.

[39] Dias, G. H., Hasanuzzaman, M., Ferrari, S., and Mathet, Y. Tempowordnet for sentence time tagging. In *Proceedings of the 23rd International Conference on World Wide Web* (2014), ACM, pp. 833–838.

[40] Ding, X., Liu, B., and Yu, P. S. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (2008), ACM, pp. 231–240.

[41] Ding, X., Liu, B., and Zhang, L. Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), ACM, pp. 1125–1134.

[42] Ebrahimi, M., Yazdavar, A. H., and Sheth, A. Challenges of sentiment analysis for dynamic events. *IEEE Intelligent Systems 32*, 5 (2017), 70–75.

[43] Esuli, A., and Sebastiani, F. Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation* (2007), 1–26.

[44] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web* (2004), ACM, pp. 100–110.

[45] Finn, C., and Lizier, J. T. Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy 20*, 4 (2018), 297.

[46] Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (2003), Association for Computational Linguistics, pp. 168–171.

[47] Fukuhara, T., Nakagawa, H., and Nishida, T. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In *ICWSM* (2007).

[48] Ganesan, K., and Zhai, C. Opinion-based entity ranking. *Information retrieval 15*, 2 (2012), 116–150.

[49] Gurini, D. F., Gasparetti, F., Micarelli, A., and Sansonetti, G. Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization. *Future Generation Computer Systems 78* (2018), 430–439.

[50] Haddi, E., Liu, X., and Shi, Y. The role of text pre-processing in sentiment analysis. *Procedia Computer Science 17* (2013), 26–32.

[51] Han, B., and Lavie, A. A framework for resolution of time in natural language. *ACM Transactions on Asian Language Information Processing (TALIP) 3*, 1 (2004), 11–32.

[52] Hatzivassiloglou, V., and McKeown, K. R. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics* (1997), Association for Computational Linguistics, pp. 174–181.

[53] Hatzivassiloglou, V., and Wiebe, J. M. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1* (2000), Association for Computational Linguistics, pp. 299–305.

[54] Hung, C., Tsai, C.-F., and Huang, H. Extracting word-of-mouth sentiments via sentiwordnet for document quality classification. *Recent Patents on*

Computer Science 5, 2 (2012), 145–152.

[55] Jindal, N., and Liu, B. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (2008), ACM, pp. 219–230.

[56] Joshi, A., Balamurali, A., and Bhattacharyya, P. A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON* (2010).

[57] Kaur, A., and Gupta, V. Proposed algorithm of sentiment analysis for punjabi text. *Journal of Emerging Technologies in Web Intelligence 6*, 2 (2014), 180–183.

[58] Khan, F. H., Bashir, S., and Qamar, U. Tom: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems 57* (2014), 245–257.

[59] Khan, F. H., Qamar, U., and Bashir, S. Sentimi: Introducing point-wise mutual information with sentiwordnet to improve sentiment polarity detection. *Applied Soft Computing 39* (2016), 140–153.

[60] Kothapalli, M., Sharifahmadian, E., and Shih, L. Data mining of social media for analysis of product review. *International Journal of Computer Applications 156*, 12 (2016).

[61] Kucuktunc, O., Cambazoglu, B. B., Weber, I., and Ferhatosmanoglu, H. A large-scale sentiment analysis for yahoo! answers. In *Proceedings of the fifth ACM international conference on Web search and data mining* (2012), ACM, pp. 633–642.

[62] Larsen, B., and Aone, C. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (1999), ACM, pp. 16–22.

[63] Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., and Lauw, H. W. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (2010), ACM, pp. 939–948.

[64] Liu, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies 5*, 1 (2012), 1–167.

[65] Liu, L., Kang, J., Yu, J., and Wang, Z. A comparative study on unsupervised feature selection methods for text clustering. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on* (2005), IEEE, pp. 597–601.

[66] Ljubešić, N., Erjavec, T., and Fišer, D. Standardizing tweets with character-level machine translation. In *International Conference on Intelligent Text Processing and Computational Linguistics* (2014), Springer, pp. 164–175.

[67] Lopes, L., Fernandes, P., and Vieira, R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf. *Knowledge-Based Systems 97* (2016), 237–249.

[68] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (2011), Association for Computational Linguistics, pp. 142–150.

[69] Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. Building a large annotated corpus of english: The penn treebank. *Computational linguistics 19*, 2 (1993), 313–330.

[70] McCallum, A., Nigam, K., et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (1998), vol. 752, Citeseer, pp. 41–48.

[71] Mccord, M., and Chuah, M. Spam detection on twitter using traditional classifiers. In *international conference on Autonomic and trusted computing* (2011), Springer, pp. 175–186.

[72] Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM 38*, 11 (1995), 39–41.

[73] Moghaddam, S., and Popowich, F. Opinion polarity identification through adjectives. *arXiv preprint arXiv:1011.4623* (2010).

[74] Mukherjee, A., Liu, B., and Glance, N. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web* (2012), ACM, pp. 191–200.

[75] Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., and Pham, S. B. A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. *AI Communications 29*, 3 (2016), 409–422.

[76] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM 11*, 122-129 (2010), 1–2.

[77] Pandey, P., and Govilkar, S. A framework for sentiment analysis in hindi using hswn. *International Journal of Computer Applications 119*, 19 (2015).

[78] Pang, B., and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (2004), Association for Computational Linguistics, p. 271.

[79] Pang, B., Lee, L., et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval 2*, 1–2 (2008), 1–135.

[80] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (2002), Association for Computational Linguistics, pp. 79–86.

[81] Pennell, D. L., and Liu, Y. Evaluating the effect of normalizing informal text on tts output. In *Spoken Language Technology Workshop (SLT), 2012 IEEE* (2012), IEEE, pp. 479–483.

[82] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.-S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (2016), pp. 19–30.

[83] Razavi, A. H., Matwin, S., De Koninck, J., and Amini, R. R. Dream sentiment analysis using second order soft co-occurrences (sosco) and time course representations. *Journal of Intelligent Information Systems 42*, 3 (2014), 393–413.

[84] Renduchintala, A., Knowles, R., Koehn, P., and Eisner, J. Creating interactive macaronic interfaces for language learning. *ACL 2016* (2016), 133.

[85] Riloff, E., and Wiebe, J. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural*

language processing (2003), Association for Computational Linguistics, pp. 105–112.

[86] Sharma, R., and Bhattacharyya, P. A sentiment analyzer for hindi using hindi senti lexicon. In *11th International Conference on Natural Language Processing* (2014), p. 150.

[87] Sharma, R., Nigam, S., and Jain, R. Polarity detection movie reviews in hindi language. *arXiv preprint arXiv:1409.3942* (2014).

[88] Srivastava, R., Bhatia, M., Srivastava, H. K., and Sahu, C. Exploiting grammatical dependencies for fine-grained opinion mining. In *Computer and communication technology (iccct), 2010 international conference on* (2010), IEEE, pp. 768–775.

[89] Strapparava, C., and Mihalcea, R. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (2008), ACM, pp. 1556–1560.

[90] Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment in twitter events. *Journal of the Association for Information Science and Technology 62*, 2 (2011), 406–418.

[91] Thelwall, M., Buckley, K., and Paltoglou, G. Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology 63*, 1 (2012), 163–173.

[92] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*, 12 (2010), 2544–2558.

[93] Tsai, A. C.-R., Wu, C.-E., Tsai, R. T.-H., and Hsu, J. Y.-j. Building a concept-level sentiment dictionary based on commonsense knowledge. *IEEE Intelligent Systems 28*, 2 (2013), 22–30.

[94] Turney, P. D. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *European Conference on Machine Learning* (2001), Springer, pp. 491–502.

[95] Turney, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (2002), Association for Computational Linguistics, pp. 417–424.

[96] Wang, G., Xie, S., Liu, B., and Yu, P. S. Identify online store review spammers via social review graph. *ACM Transactions on Intelligent Systems and Technology (TIST) 3*, 4 (2012), 61.

[97] Wang, J.-z., Yan, Z., Yang, L. T., and Huang, B.-x. An approach to rank reviews by fusing and mining opinions based on review pertinence. *Information fusion 23* (2015), 3–15.

[98] Wilson, T., Wiebe, J., and Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on*

human language technology and empirical methods in natural language processing (2005), Association for Computational Linguistics, pp. 347–354.

[99] Wu, D. D., Zheng, L., and Olson, D. L. A decision support approach for online stock forum sentiment analysis. *IEEE Transactions on systems, man, and cybernetics: systems 44*, 8 (2014), 1077–1087.

[100] Xie, S.-x., and Wang, T. Construction of unsupervised sentiment classifier on idioms resources. *Journal of Central South University 21*, 4 (2014), 1376–1384.

[101] Yi, J., and Niblack, W. Sentiment mining in webfountain. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on* (2005), IEEE, pp. 1073–1083.

[102] Zhang, T., Damerau, F., and Johnson, D. Text chunking based on a generalization of winnow. *Journal of Machine Learning Research 2*, Mar (2002), 615–637.

[103] Zhao, W., and Zhou, Y. A template-based approach to extract product features and sentiment words. In *Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on* (2009), IEEE, pp. 1–5.

[104] Zhu, J., Zhang, C., and Ma, M. Y. Multi-aspect rating inference with aspect-based segmentation. *IEEE Transactions on Affective Computing 3*, 4 (2012), 469–481.

[105] Zhu, S., Liu, Y., Liu, M., and Tian, P. Research on feature extraction from chinese text for opinion mining. In *Asian Language Processing, 2009. IALP'09. International Conference on* (2009), IEEE, pp. 7–10.

[106] Zhu, Y., and Newsam, S. Spatio-temporal sentiment hotspot detection using geotagged photos. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2016), ACM, p. 76.

Chapter 6

List of Publications

Journal Publications

1. Sukhnandan Kaur, Rajni Mohana, A Roadmap of Sentiment Analysis and its Research Directions, International Journal of Knowledge and Learning, vol.10, no.3,pp.296-323, 2015. (Scopus) (SJR-0.11)

2. Sukhnandan Kaur, Rajni Mohana,Prediction of Sentiment from Macaronic Reviews, Informatica: An International Journal of Computing and Informatics, vol.42, no. 1, pp. 127-136, 2018. (ESCI) (Scopus) (SJR-0.39)

3. Sukhnandan Kaur, Rajni Mohana, Temporality Based Sentiment Analysis using Linguistic Rules and Meta-Data, Proceedings of the National Academy of Sciences, India Section A: Physical Sciences , 2018.

4. Akanksha Puri, Sukhnandan Kaur, Rajni Mohana, Temporal Sentiment Analysis: A Review, International Journal of Control Theory and Applications,vol.9, Issue.40, pp.327-334, 2016.

Conference Publications / Book Chapters

1. Sukhnandan Kaur, Rajni Mohana, Unsupervised Document Level Sentiment

Analysis of Reviews using Macaronic Parser, in proceedings of Forth International Conference on Emerging Research in Computing, Information, Communication and Applications, ERCICA, Springer, Banglore, India, July, 2016.

2. Sukhnandan Kaur, Rajni Mohana, Prediction of Sentiment from Textual Data Using Logistic Regression Based on Stop Word Filteration and Volume of Data, Shannon 100, Jalandhar, India, April, 2016.

3. Ashima, Sukhnandan Kaur, Rajni Mohana, Anaphora Resolution in Hindi: A Hybrid Approach, Intelligent Systems Technologies and Applications, vol. 530, Springer, pp. 815-830, 2016.

Journal Publications

1. Sukhnandan Kaur, Rajni Mohana, A Roadmap of Sentiment Analysis and its Research Directions, International Journal of Knowledge and Learning, vol.10, no.3,pp.296-323, 2015. (Scopus) (SJR-0.11)

2. Sukhnandan Kaur, Rajni Mohana,Prediction of Sentiment from Macaronic Reviews, Informatica: An International Journal of Computing and Informatics, vol.42, no. 1, pp. 127-136, 2018. (ESCI) (Scopus) (SJR-0.39)

3. Sukhnandan Kaur, Rajni Mohana, Temporality Based Sentiment Analysis using Linguistic Rules and Meta-Data, Proceedings of the National Academy of Sciences, India Section A: Physical Sciences , 2018.

4. Akanksha Puri, Sukhnandan Kaur, Rajni Mohana, Temporal Sentiment Analysis: A Review, International Journal of Control Theory and Applications,vol.9, Issue.40, pp.327-334, 2016.

Conference Publications / Book Chapters

1. Sukhnandan Kaur, Rajni Mohana, Unsupervised Document Level Sentiment Analysis of Reviews using Macaronic Parser, in proceedings of Forth International Conference on Emerging Research in Computing, Information, Communication and Applications, ERCICA, Springer, Banglore, India, July, 2016.

2. Sukhnandan Kaur, Rajni Mohana, Prediction of Sentiment from Textual Data

Using Logistic Regression Based on Stop Word Filteration and Volume of Data, Shannon 100, Jalandhar, India, April, 2016.

3. Ashima, Sukhnandan Kaur, Rajni Mohana, Anaphora Resolution in Hindi: A Hybrid Approach, Intelligent Systems Technologies and Applications, vol. 530, Springer, pp. 815-830, 2016.