

HOLISTIC MULTILINGUAL SENTIMENT ANALYSIS ON REVIEWS IN SOCIAL MEDIA

Synopsis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE & ENGINEERING

By

SUKHNANDAN KAUR

Enrollment No. 136215



Under the supervision of

DR. RAJNI MOHANA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
WAKNAGHAT

Table of Contents

	Page Number
Abstract.....	i
1. INTRODUCTION.....	1
2.LITERATURE REVIEW.....	5
3. CONTRIBUTION OF THE WORK	7
3.1. Normalization of the web content for sentiment analysis	7
3.2. Handling macaronic text	12
3.3. TempoSentiscore generation for sentiment analysis	18
3.4. Analysis of various supervised learning approaches	21
3.5. Conclusion and future scope	23
References	26
List of Publications	33

ABSTRACT

Various natural language processing (NLP) tasks are carried out to feed into computerised decision support systems. Among these, sentiment analysis is gaining more attention. Sentiment analyzers are highly useful in enterprise business also. These systems aim to aid decision making for customers, manufacturers, etc. by providing easily accessible information when needed. There are huge number of social media sites such as Twitter, Facebook, BlogSpot, Amazon, etc. which are used for collecting the reviews of people about any entity. The web users act as an advisory body for various enterprises. Business people use this data for figuring out the major and minor flaws in their products or services. The primary 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. Along with this, methods for effective pre-processing are also introduced i.e. textual normalization.

The majority of state-of-the-art approaches to sentiment analysis rely on the social media content. Due to the growth in social media, the number of words per post is limited which give rise to the use of short hands, slangs, etc. This web content is highly un-normalized in nature. This hinders the performance of decision support system. To enhance the performance of decision support system, it is very much required to process data efficiently. This rises to formulate the novel method for textual normalization of web data during pre-processing phase which includes handling emoticons. In this, we have used semantic mapping of emoticons. We have used hybrid method which comprises two basic modules: cross-word dictionary and corpus based approach. It is aimed to get better results for different natural language processing tasks for automated decision support systems.

Nowadays, there is an exponential rise in the available internet data. People prefer writing in native language either full content or some of the content. For handling full content, various multilingual approaches are used. The problem is how to deal when some foreign language content is studded into base language content. This helps us to formulate such a system to deal with this type of content i.e. macaronic content. Proposed algorithm outperforms state of the art sentiment analysers which simply discard such content.

Outdated 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. For the generation of temposentiscore linguistic rules

as well as meta data associated with the content has considered. Temposentiscore results have been compared with sentisore generated by twitter opinion mining(TOM) algorithm. Effectiveness of proposed temposentiscore generation of web data is demonstrated with the help of star rating.

Finally, we have analysed various learning algorithms based on different performance metrics. Effect of proposed pre-processing i.e. textual normalization has been analysed

1. INTRODUCTION

The intersection between social media and user generated content arose a great deal of research in the area of sentiment analysis (SA). SA is present in many spheres of our daily lives, whether we realise or not. It affects how we shop, work, sale, etc. SA is a collaborative process of natural language processing and data mining. The work in SA is a subset of text engineering. This can also be defined as to process various sentiment signals to support automatic decision generation. Diffusion of sentiment signals into binary form is the prime task of SA, i.e. positive and negative. Refinement in the level of granularity was proliferated along with the technical hikes in the machine learning.

1.1 Evolution in Sentiment Analysis^[1]

During the time span of early 2000s, researchers were working on the polarity check on the document, i.e., positive, negative or neutral signals. Prior, the evaluation was done at the document level, but gradually drifted to sentence level (i.e., considering only subjective sentences) and nowadays entity/feature level is increasing. Due to this evolution, definition of SA has also changed it as shown in Fig. 1.

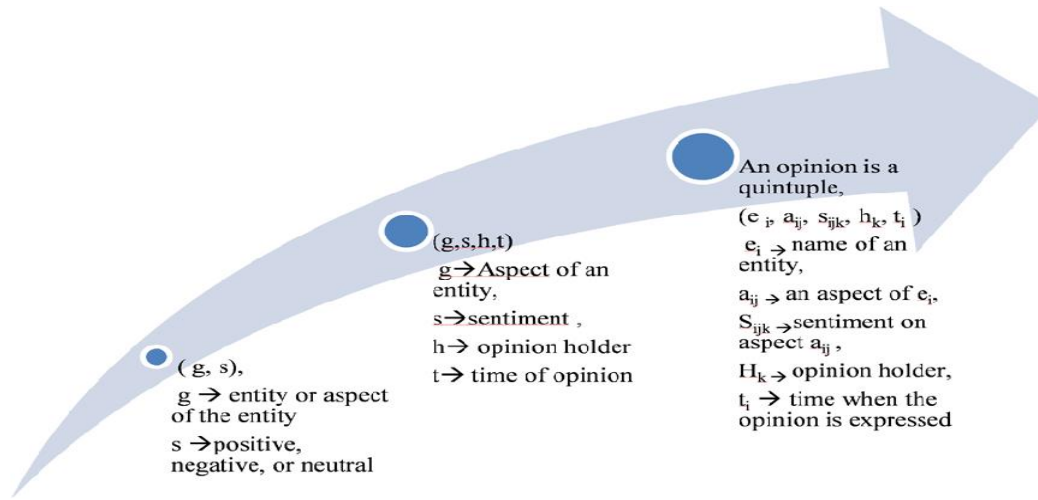


Fig. 1: Evolution of Sentiment Analysis

Definition 1^[1]: A SA is a process having binary tuples, ‘e’ is the entity for which the document is about and ‘s’ is the sentiment about the document, i.e.,

$$SA = \{e, g\}.$$

It gives the output as the only entity with its corresponding opinion. This definition does not focus on the issue that “*Who is the opinion holder*”. The opinion holder may also change his opinion with time, due to which the dissimilarity can arise. Thus, researchers came up with another definition:

Definition 2^[1]: SA composed of four tuples, ‘ g ’ is an aspect of the entity, ‘ s ’ is sentiment, ‘ h ’ is opinion holder and ‘ t ’ is time of opinion, i.e., $SA = \{g, s, h, t\}$. It was realised that a document can contain views about more than one aspect of an entity. To cope up with this problem definition was again revised.

Definition 3^[1]: SA now 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 and ‘ t ’ is time of opinion, i.e., $SA = \{e, a, s, h, t\}$.

These evolution makes sentiment analysis refined using finer level of granularity. In this thesis, we focus on the temporality, unstructured web data in textual form and multilingual content.

1.2. Research Gaps

By reviewing the history, it tends to form a need of deeply analyse the existing research work. Various research aspects which are still untouched or needs more attention. There are many available survey in SA but in this thesis, we attempt to compile the research done till date. We also have identified various research directions-like unstructured sentences, temporal tagging and multi-lingualism.

1.2.1. Temporality

With coming technological era, people are much aware of the importance of communication. Delivering right message at the right time gives good effect over others. If the manufacturer know about the flaws in his product or able to know the criticism of the people towards the product, he might be able to deal with it at the right time. It not only give him good results but also swing the mood of people towards his product. Sentiment analysis gives best results if temporality is captured along with.

1.2.2 Normalization

People who triggered the tweets or any content over social media hardly think much about the syntactic and semantic structure of the content. They sometimes use slangs, emoticons or their native language words to blow out their sentiment about any entity. This makes the task of data analysis complex. This sometimes contains less or more noisy data. So, before analyzing the sentiments attached to that content, we firstly pre-process the data. This not only gives us effective results but also increase the reliability of decision support system.

1.2.3 Multilinguality

In this multilingual heterogeneous web content, different societies use different languages and their way of writing is also varies. They have the freedom to use their native language too. Due to the scarcity of the language resources over the web, it becomes very difficult to handle all the possible language over the globe. It is a challenging task of a natural language processing. The irregularities found in the data over the internet make it more complex. Depends on the availability of the majority group in a society, nation, nation-state, or community, monolingual systems are designed. These systems somehow have very limited number of formal users. This type of formalism in sentiment analysis limit the system to specific users. The reviews from all the users of a particular entity are valuable. A multilingualism in sentiment analysis raises the need for having an ideal sentiment analysers which makes the user free to write anything .i.e. multilingual sentiment analysis. People these days write their reviews in compound language words. The reviews from all the users should be included for sentistrength detection.

1.3. Performance measurement^[1]

There are the following metrics broadly used to evaluate the system performance:

Precision, Recall, F-measure, human annotator agreement and accuracy.

To calculate precision, recall and accuracy the following metrics need to be defined:

- *True positives (TP)*: Number of positive examples labelled as positive.
- *False positives (FP)*: Number of negative examples labelled as positive.
- *True negatives (TN)*: Number of negative examples labelled as negative.
- *False negatives (FN)*: Number of positive examples labelled as negative.
- *Correct output (CO)*: Number of outputs of the system which are considered as correct by the human annotators.

1. *Recall* : It is the percentage of named entities present in the corpus that are found by the learning system. It is poor in case of less training data due to which the system is unable to cover all the terms.

$$\text{Recall}(R) = TP / (TP + FN)$$

2. *Precision* : It is the accurate number of named entities found by the learning system. It is found high if it gives correct results.

$$\text{Precision}(P) = TP / (TP + FP)$$

3. *Accuracy*: It is defined as the ratio of addition of true positive, true negative and true positive, true negative, false positive, false negative.

$$\text{Accuracy}(A) = (TP + TN) / (TP + TN + FP + FN)$$

4. *Human agreement (HA)*: It is the percentage of agreement between two or more humans for the correct output of the system. It is measured in terms of accuracy of the system by human annotators. It is also called as human annotator agreement.

$$\text{Human annotator agreement} = (CO) / (TP + TN + FP + FN)$$

1.4. Dataset

i) Polarity dataset v1.1^[1].

Consists 700 reviews of positive and 700 for negative from the benchmark dataset vol.1.

Consists 1k reviews of positive and 1k for negative from the benchmark dataset vol.2.

ii) Customized data for Macaronic language.

Dataset with manual subjective annotations using inter annotator agreement.

It consists 1k reviews in positive and 1k in negative category to remove biasing.

iii) Reviews for different models of Ford car^[3].

Total number of reviews: ~42,230

~ 18,903 reviews in each category

1.5. Hardware

We ran the algorithms on Window 7 32-bit operating system IntelA.

i3 processor 530@ 2.93 GHz machine.

1.6. Objectives

The main contribution of the thesis is explained as:

- 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. LITERATURE REVIEW

2.1. *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 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 value for positive and negative terms appeared in any document. Pang et al.[2] used fine grained corpus for not only detecting the sentiment but also the implicit aspect and the global entity about which the sentiment is generated. They have used camera and mobile phone reviews for their work. This also enhanced the work of implicit entity detection in sentiment analysis. In their work they have shown the importance of corpus based sentiment analysis. 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. Hogenboom et al. [3] 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. Outperformance of the results of their system with respect to the state of art sentiment analysers was shown in their work. Lin et al.[4] have identified the circumstances for the growth of individual advancement for cross-disciplinary work. Dictionary based approach was included. In their work, they have primarily embedded artificial intelligence in the form of neural networks for opinion mining. Different agent based approaches are employed for their work. Their work results are significant in some areas but are 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. McCord et al.[5] used random walk and iterative regression for first building concept level lexicon. They have used commonsense for the annotation of the lexicons. Later, the correlation between stock price and the reviews of stock finance through sentiment analysis was found. Their work highlights the need of accessing the reviews efficiently for automation of decision support system. Yuhai et al.[6] developed a strategy to extract sentiment from textual as well as visual web data in a combined way. The results are better as compared to state-of-art sentiment analysis in Chinese language as well as visual sentiment analysis individually. From the recent studies, we have analyzed that some researchers used dictionary based technique and some used corpus based approach to deal with the slangs, misspelled words, emoticons, etc. This paper proposes a novel hybrid approach which uses lexicon and corpus based approach in a combined manner. The embedding of both lexicon and corpus based approach gives better results for sentiment analysis.

2.2 *Normalization*

Nikola et al.[7] have normalized the text using character-level statistical machine translation system and training through manually annotated dataset. From their work, it is 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 Francois et al. [8]. They have worked over the normalization of the textual data using finite state transducers. Results have 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. Alistair et al.[9] have used the variations in spellings written by people. 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, Deana et al.[10] has proved the effectiveness of normalized text over the performance of text to speech system. Their research also helped normalization in gaining more attention. Wang et al.[11] 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 translator's effectiveness. Tyler et al.[12] said in their work that normalization was not the matter of just replacing the words, but it actually depends on the target application. System was designed by them to handle domain dependent normalization. Zhang et al.[13] worked for performing all the necessary steps to have the formal structured content. Chieu et al.[14] have used the corpus based technique for short message normalization. In their work, translation of the short messages to formal English language was handled. Chieu et al.[15] and Nasukawa et al. [16] both normalized the text independent to the discipline in which it was used. Our approach is also based on this fact to have a normalized text for general purpose.

2.3 Multilinguality

Earliest works in the area of sentiment analysis is done by Hatzivassiloglou et.al[17]. They have used adjectives for deducing the polarity of the document. His work is then elaborated by Pang et.al.[18]. They mainly focused over the supervised learning algorithms. They have used various machine learning algorithms for their work. The continuation of their work in the field of sentiment analysis is given in [19,20], where they have used minicut algorithm for opinion summarization and also presented various opinion mining techniques. Latest research paying more attention to the sentiment analysis over the data collected from various social sites. Researchers [21,22] have used social media data for the sentiment analysis. Connor et.al. [22] have used twitter data for sentiment analysis based on time series. Yang and Liang [23] proposed a new approach for identification of natural language, i.e., joint approach based on N-Gram frequency statistics and Wikipedia. Carter et al.[24] have used N-gram approach to identify five languages from the microblog. Later Lui et.al.[25] followed same approach over the long and short textual documents. A detailed study of sentiment analysis at various levels of granularity[27] is explained by Kaur et.al. Out of this literature study, we have found that the inclusion of normalization of macaronic language still needs attention. In this manuscript, we proposed a model to deal with the macaronic language for sentiment analysis.

CONTRIBUTION OF THE WORK

The contribution of the thesis is based on the following objectives formulated from the research gaps. The objectives of the thesis are normalization of the web content including handling emoticons as well as macaronic content, capturing the essence of temporality and analyse various machine learning algorithms.

3.1. NORMALIZATION OF THE WEB CONTENT FOR SENTIMENT ANALYSIS

3.1.1 Introduction

Natural language processing 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. In this process, initially, data is collected from the various social sites for automation of the decision support systems. Then data is pre-processed to get the structured content 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 is 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 people's 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 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 is 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 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.

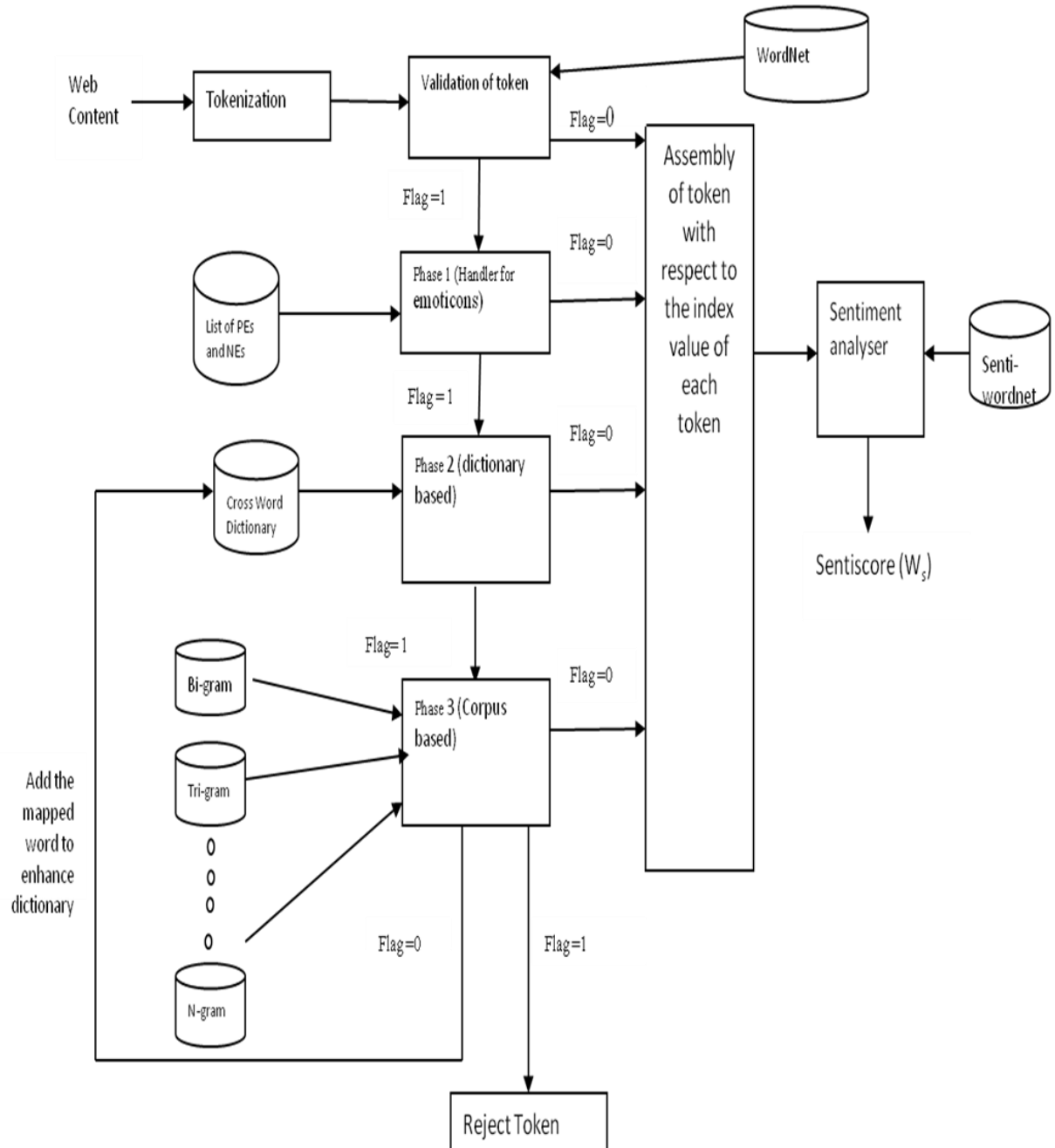


Fig. 2: Proposed System Design for Normalization

3.1.2. Approach

Our proposed work includes the system design and algorithm to handle unstructured or noisy data for sentiment analysis. Although the presented algorithm is generic in nature, we have applied and tested it for sentiment analysis. The hybrid framework of the proposed system for sentiment analysis primarily consist pre-processing and sentiment strength calculation as shown in Fig.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 pre-processing 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. We have normalized the data by eliminating the slangs, emoticons, noisy text, etc using dictionary.

3.1.3. Results and discussion

On comparison from table 1, we found better results for the normalized dataset than un-normalized data. This shows the importance of normalization for any natural language processing task. Another significance of normalized data is reduces the fallout. For Naïve-Bayes, we get equal fallout for normalized and un-normalized dataset. On the contrary, we have also found the recall and precision values are raised in normalized data processing. Fig.3 shows the performance of various supervised and unsupervised approaches. We found better results using SVM. It is noticeable that there is a trade-off between recall and fallout, precision and fallout. Naïve Bayes gives minimum fallout and k-NN has maximum value of recall, precision and accuracy. The highest value of fallout in k-NN reduces its performance in some cases wherever we need less fallout value. It depends on the application and domain to choose the learning algorithm.

After successful normalization using hybrid approach, we have applied sentiment analysis over the given dataset. We have shown the variations in the results using Fig.3. It clearly shows that the normalization affects the results of sentiment analyzer to a great deal. After applying hybrid normalization on the dataset for sentiment analysis, we have found variations in the documents fall in positive and negative category shown in Fig. 4. These results may

differ for different datasets. Results getting after normalization are more realistic as it is more close to the gold standard shown in Fig. 5. We have found that hybrid approach based normalization gives better results than rule based and corpus based in terms of average accuracy.

Table 1: Results for both un-normalized and normalized datasets.

Dataset	Model	Precision	Recall	Accuracy	F-measure	Fallout
Un-normalized	SVM	51.38	50.91	73.33	51.14	59.18
	NB	55.69	56.14	68.57	55.91	02.68
	k-NN	66.95	66.14	79.55	66.79	24.77
Normalized	SVM	53.29	52.05	75.16	52.66	59.18
	NB	59.93	60.68	73.44	60.33	02.68
	k-NN	66.74	64.91	79.55	65.81	40.00

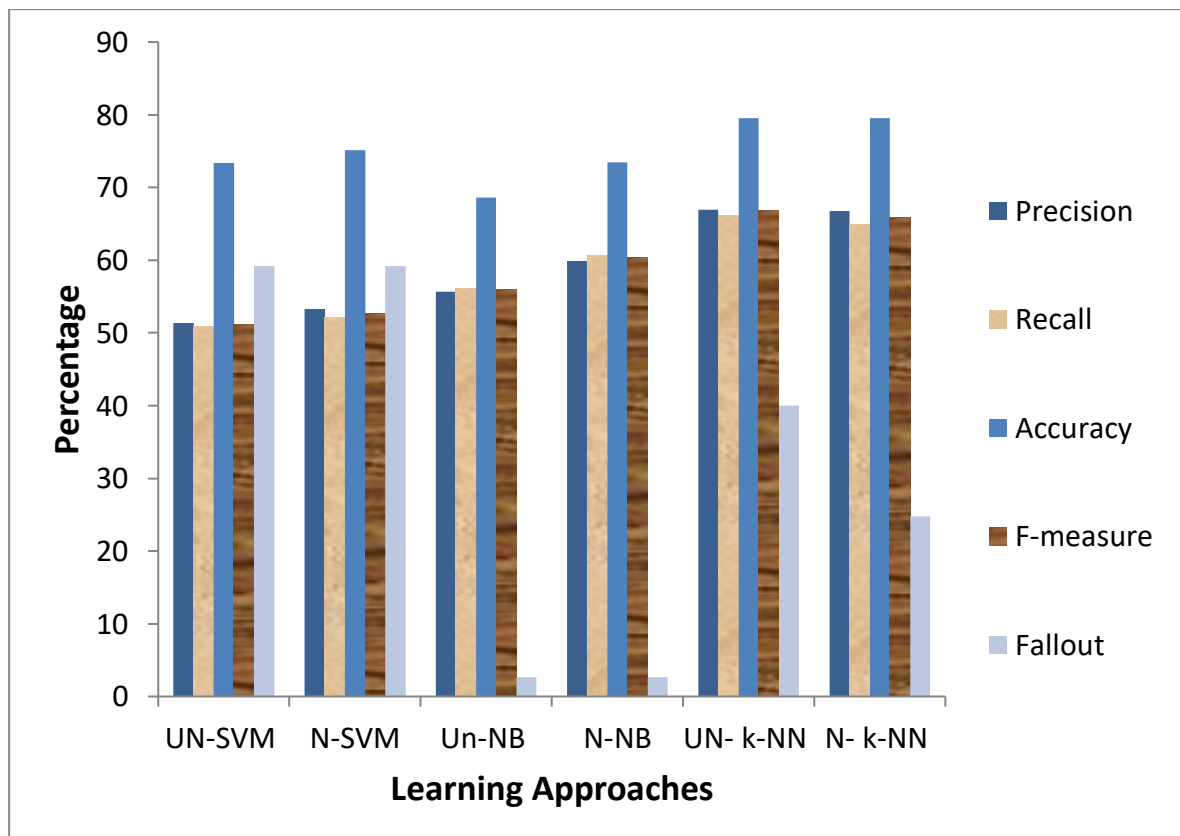


Fig. 3: Results Based On Normalization

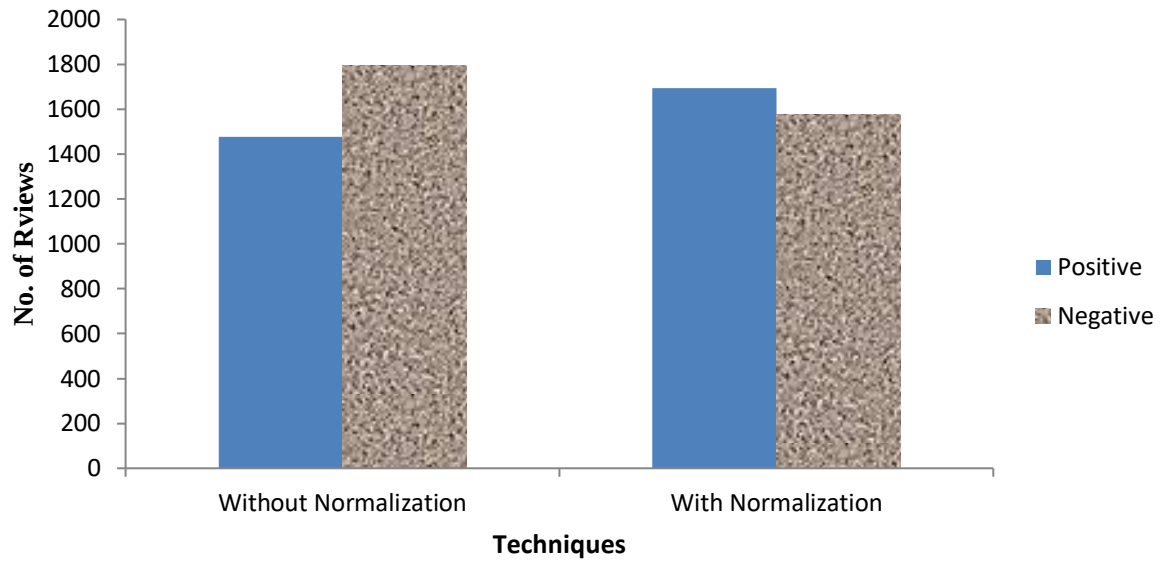


Fig. 4: Categorization of Reviews based on Proposed Algorithm

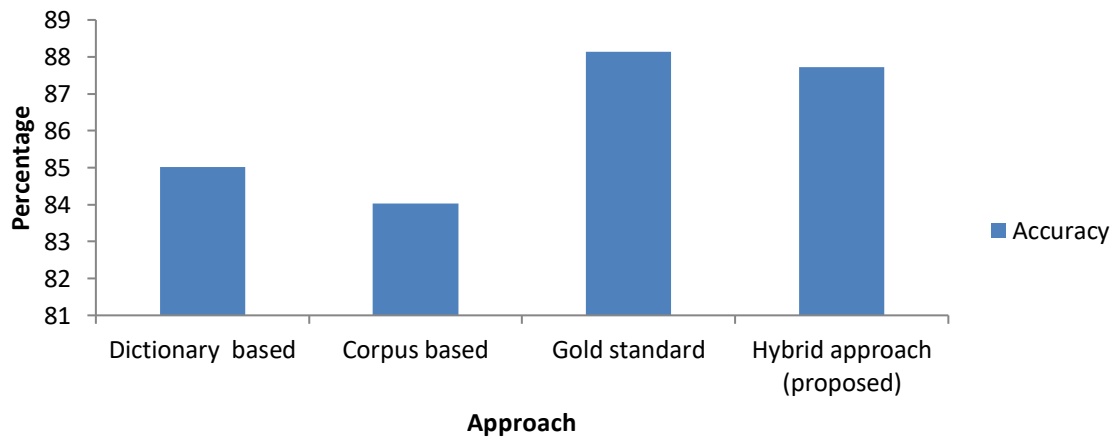


Fig.5: Comparison with Gold Standard

Results getting after normalization are more realistic as it is more close to the gold standard shown in Fig.5. We have found that hybrid approach based normalization gives better results than rule based and corpus based in terms of average accuracy.

3.1.4.Conclusion

To conclude, we can say that the proposed algorithm is very close to the Gold standard. Therefore, it is very effective in case of sentiment analysis. It refines the positive and negative category reviews. Due to its generic nature, it can be applied to other domains also. The

generality of the algorithm makes it beneficial in various language processing tasks such as summarization, named entity recognition, etc.

3.2. HANDLING MACARONIC TEXT

3.2.1.Introduction

Sentiment analyzers are highly useful in enterprise business. There are huge number of social media sites such as Twitter, Facebook, BlogSpot, Amazon, etc. which are used for collecting the reviews of people about any entity. The web users act as an advisory body for various enterprises. Business people use this data for figuring out the major and minor flaws in their products or services. This also helps them to improve their product quality. There is no language barrier to write anything over the internet. This makes the task of sentiment analyzers a bit complex. Nowadays, there is large number of sentiment analysers available for different languages. Along with multilingual text, people also use macaronic language over the internet. Basically macaronic language consists multilingual text which comprises of different languages/scripts together. With growing diversity, it has become of utmost importance that we acknowledge the existence of this kind of text. Especially, in the world where expressing opinions from anywhere in the world has become a fairly easy task. Twitter and Facebook information is the best way to keep a tab on the ongoing events, opinion of general public, trending topics etc. However, one big challenge of this kind of information mining is the redundant and incongruous elements, we find along the way. Handling macaronic language not only useful in sentiment analysis but also in many natural language processing tasks such as named entity reorganization , pronoun resolution, feature extraction, etc.

However, there stands an obstacle in our way, while mining the text in one language; we seldom are able to handle a different language in the same context. Present day analysers generally treat the other language/script words as foreign words, and lose major information in not treating these words. Processing this information is very useful for various automatic language processing tasks i.e. named entity recognition, pronoun resolution, and automatic summarization along with sentiment analysis.

The given sentence is an example of macaronic text, it consists words other than base language.

Example1: I am feeling अच्छा by watching this movie.

Here the Hindi language word (अच्छा) of the context is taken as garbage and tossed aside and the English portion of it will be taken into consideration. With this, meaningful information is lost. Here, the opinion about the movie is missed. We don't particularly know the opinion because it has been tossed aside. Similarly, if we apply the filter over the discarded words i.e. foreign words and converts every word from Hindi to English. We would be able to figure out the opinion about film.

Motivation

The motivation of the proposed technique is to handle the macaronic text by automatic identification of the fragments of the text belongs to different languages. The existing systems often discard the words other than the base language. The processing of the raw data often takes the text in multiple languages as an input. Sometimes, it discards text containing meaningful information. The proposed technique is designed to handle this type of discarded fragments. From example 1, it can be clearly seen that how important is the need to normalize the macaronic text for sentiment analysis. The state of the art sentiment analyzers gives the neutral opinion about the movie although it is positive.

This arose the need to normalize the macaronic text. The proposed method is to fragment the text and auto-detect the language used in the text based on Unicode information at a script level which is different for every language/script. The deduction of the corresponding language of the specific fragment other than the target language is also presented. Hence, translate particular foreign text into a base language. For our convenience, we have taken the English as a base language.

3.2.2. Approach

The following approach has been taken to isolate and identify the language before embarking up on the journey to neutralize it through translation. The system design for the macaronic parser is given in Fig.6. Processing of the text passes through different phases of the system.

Phase 1: In this phase, the input is given to the system as a web content i.e. macaronic text/simple.

Phase 2: Based on tokens, filtration of the text is done at this level. The division of the text into base language and foreign words i.e. other than the base language is done. The tokens which are from the base language are separated first along with their index values. The foreign words are then actually processed. These foreign words are tagged as non English tokens. We have used the English as base language for our work. Therefore, we have used English WordNet for the processing. Non-English words are encoded as UTF-8.

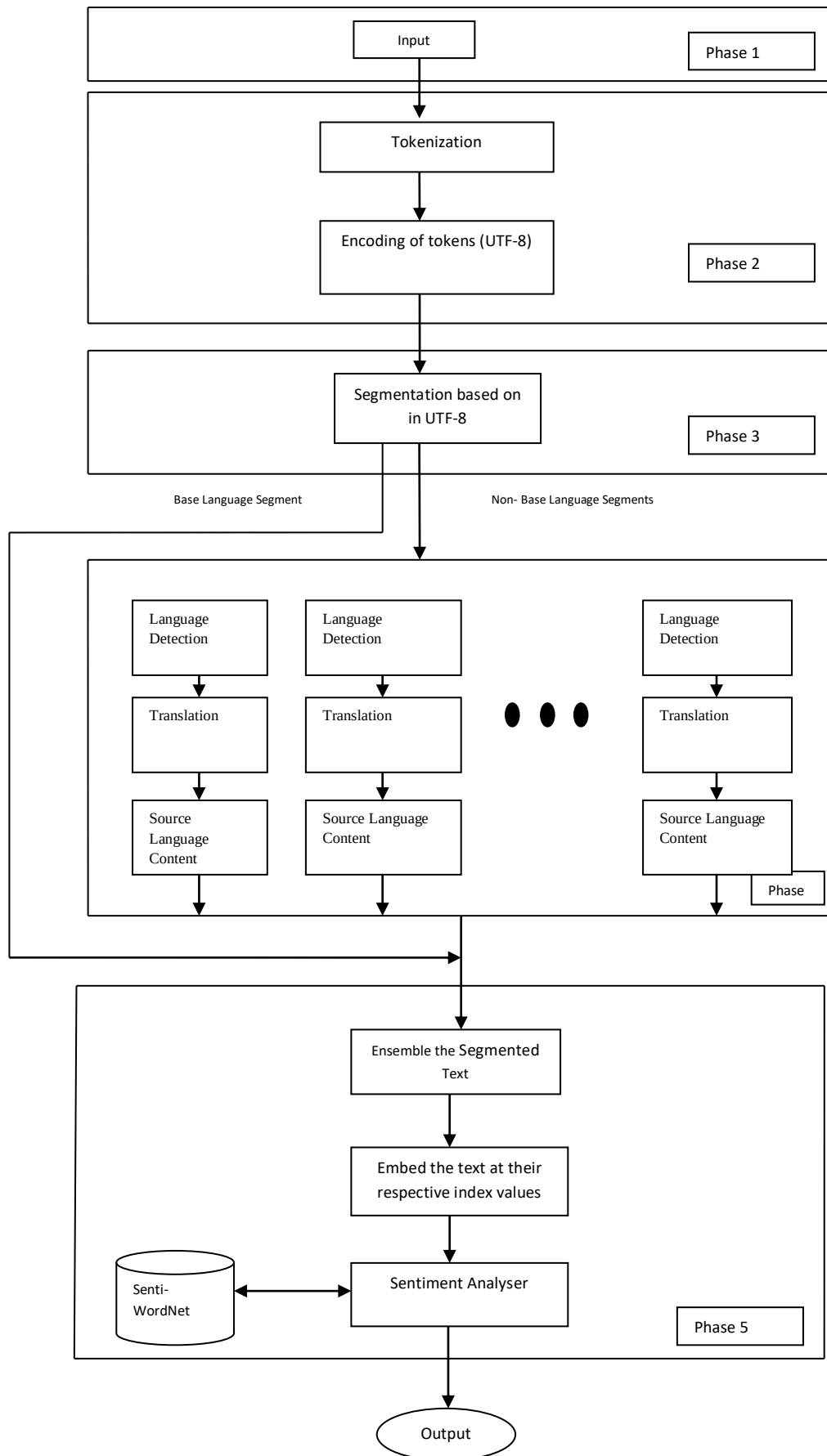


Fig.6: System Design of Macaronic Language Handler

Phase3: In this phase, segmentation of the tokens is done based on UTF-8. Each token is then has its encoded value.

Phase4: Under this phase, translation or transliteration is being done depending upon the number of languages we want to handle to cope up with the macaronic language. In this manuscript, the model represents the structural format of the processing of text. Based on UTF-8 encoding, automatic language detection algorithm is applied over it. After finding the language of the token, the working of translation or transliteration is started. The output of each sub-phase is the base/source language text.

Phase5: In the final phase of the model, whole text (English/Hindi/Punjabi) is now converted into one language i.e. English. This converted text is then passed through the sentiment analyser to generate sentiscore for each document

3.2.3.Results and Discussion

We have evaluated the system over the semantically similar dataset comprises of 200 documents i.e. English dataset and the dataset consists of the Macaronic statements which have various foreign language words studded into it. We have used customed lexicon for macaronic input i.e. Hindi words in a particular English sentence. We have applied NLTK pos tagger to find the opinionated words. SentiScore associated with the compound sentence i.e. contain foreign words from the Hindi along with English is calculated. These results are divided into two categories: Normalized dataset and Un-normalized dataset. The results obtained by the proposed method are graphically shown in Fig.7, Fig.8, Fig.9 and Fig.10.

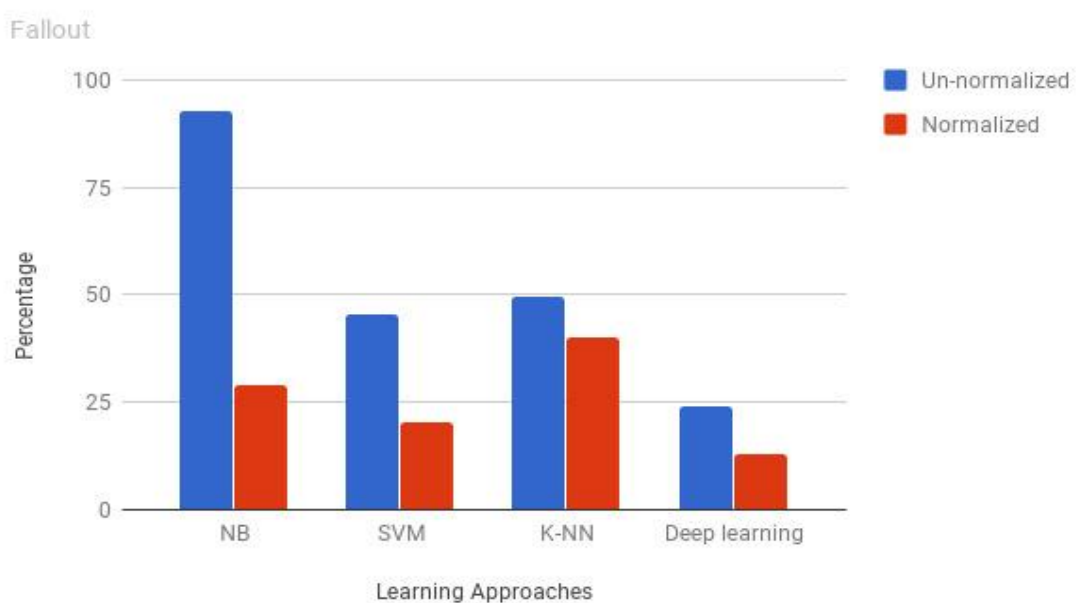


Fig. 7: Performance of Proposed System Based on Fallout

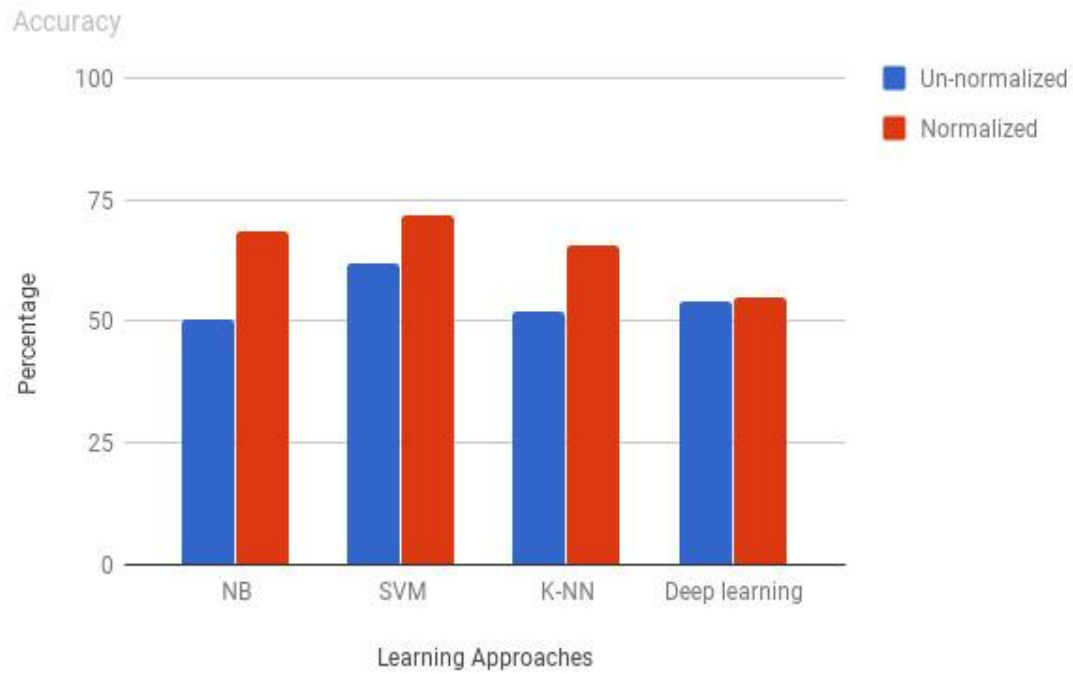


Fig. 8: Performance of Proposed System Based on Accuracy

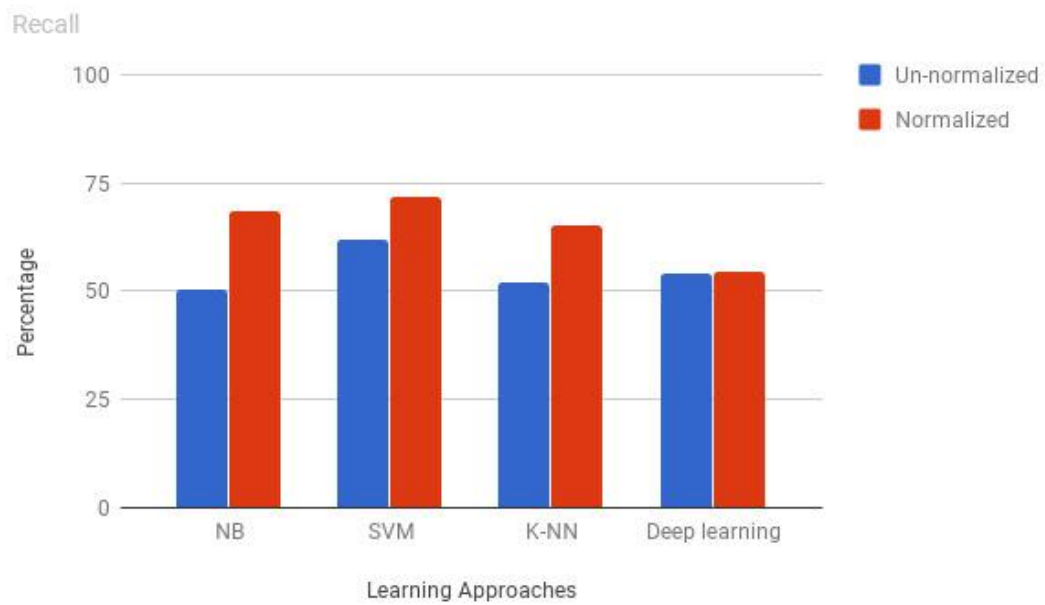


Fig. 9: Performance of Proposed System Based on Recall

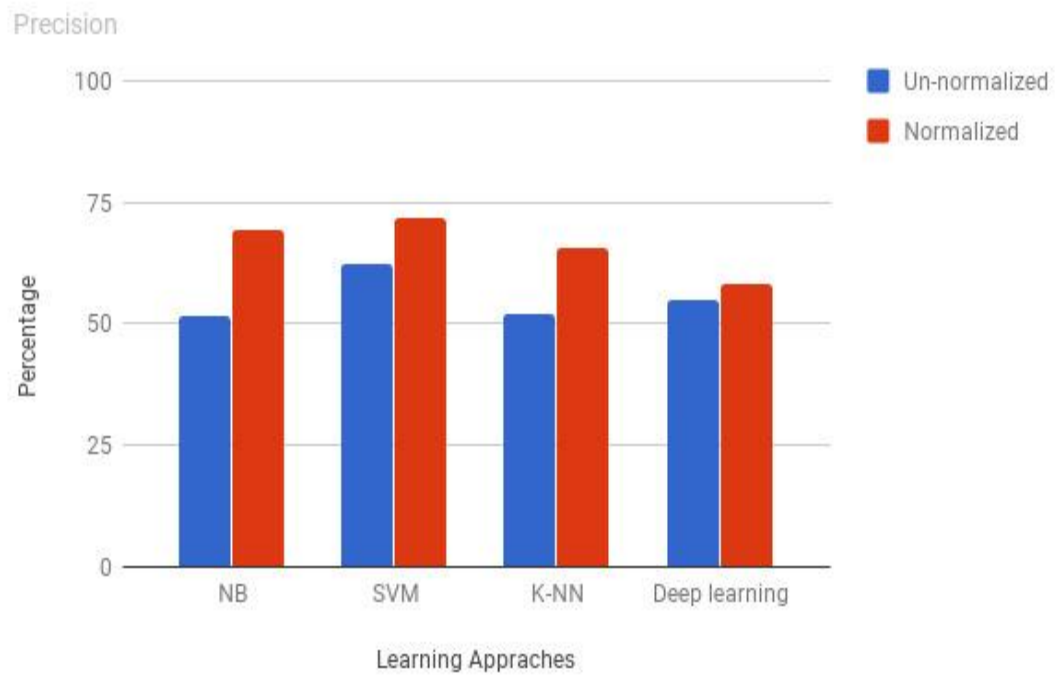


Fig. 10: Performance of Proposed System Based on Precision

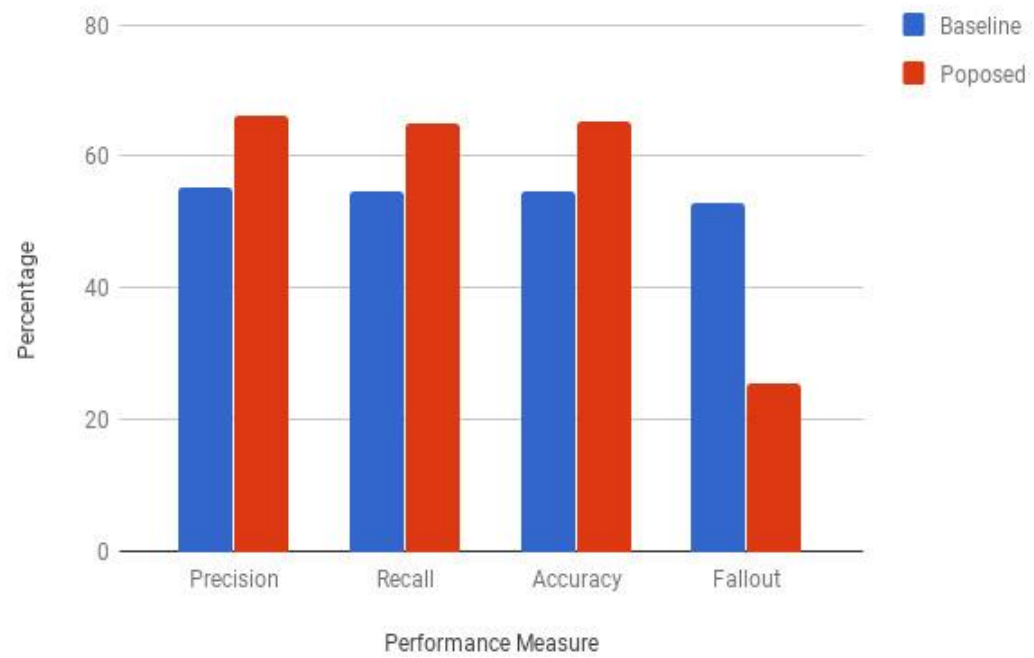


Fig. 11: Comparison of Proposed and TOM System(Baseline)

3.2.3. Conclusion

To recapitulate, it is found that the proposed system outperforms in every aspect of performance measure. Normalization greatly affects the fall-out rate in different learning approaches. Fall-out is almost reduced to half. Fig. 11 shows how proposed algorithm outperforms the Twitter Opinion Mining (baseline).

3.3. TEMPO-SENTIScore GENERATION FOR SENTIMENT ANALYSIS

3.3.1. Introduction

Outdated reviews may result in biased sentiment analysis which may or may not represent the current scenario. To remove this limitation, we are trying to implement temporal sentiment analysis of reviews by providing appropriate weightage to the reviews. 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.

Our main contribution in this thesis is the inclusion of hidden temporality i.e. Temporal tag in sentiment analysis by categorizing the documents in present, past or future, using (a) linguistic rules by employing the semantic context of words in English; (b) Meta-Data associated with each review in the dataset. The effect of these rules are shown over the standard dataset.

Another contribution is an algorithm which is designed to generate Tempo-Sentiscore. To test the performance of Tempo-Sentiscore, the Gold Standard is used. This Tempo-Sentiscore allow us to have a more precise measure of the impact of temporality on the sentiment classification task.

The objective is to better understand the impact of the temporality on the sentiment analysis. This thesis shows how the star rating get affected by considering Tempo-Sentiscore instead of Sentiscore. Through which we get the star rating closer to the real scenario.

3.3.2. Approach

To capture temporal expressions hidden in the linguistic context of words through linguistic rules as well as using metadata.

For Example: *The lens quality was good.* *Posted on: 02/05/2018*

It contains two aspects:

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

We have proposed a system that generates the sentiment analysis of the documents based on metadata as well as linguistic temporality. The architecture of the system is as shown in fig.12.

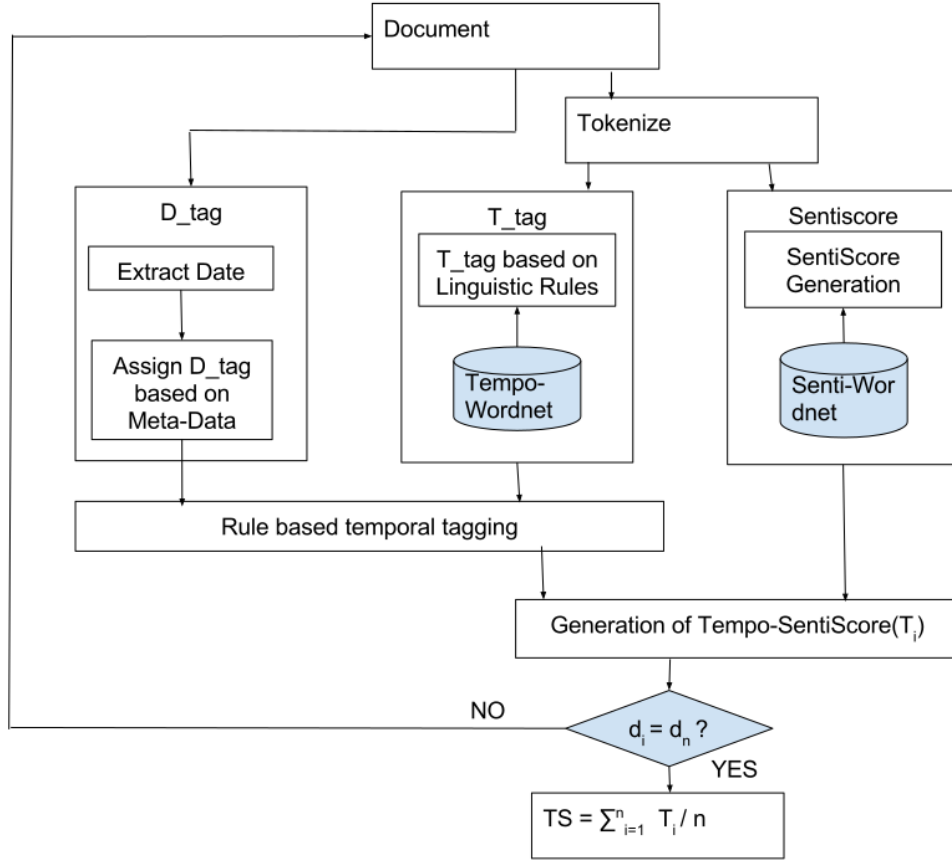


Fig. 12: System Design For Tempo-sentscore Generation

Tempo-Sentscore is calculated using equation 1. Sentscore 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.

$$TS = \frac{\sum_{i=1}^n \text{Sentscore}(\text{TempTag}_{\text{present}}) * C1 + \sum_{i=1}^n \text{Sentscore}(\text{TempTag}_{\text{past}}) * C2 + \sum_{i=1}^n \text{Sentscore}(\text{TempTag}_{\text{future}}) * C3}{n} \quad (1)$$

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 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.

Table 2: Sentiment Analysis Based on Temporal Tagging

Entity	Aggregated reviews based on temporal tag	Aggregated Sentiscore reviews	Aggregated tempo- of Sentiscore
Ford'07	past	0.57	0.19
Ford'07	Present	-0.20	-0.11
Real time Sentiscore	Past + Present	0.37	0.0895

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 to the temporal nature of the document.

Query 2: Were people happy with the ford car till 2017?

In this, the reviews till 2007 are gathered. The reviews of previous years (.....,2015,2016) are taken as past, 2017 reviews are taken as present while the reviews of 2017 pointing towards the future are taken as future .

To evaluate the effectiveness of the proposed system, we employed the experiments to overlap with the Gold standard. The variation in the Sentiscore i.e. Tempo-Sentiscore according to temporal tags is shown in Table 2.

Table 3 : Assumption of Star Rating Based on Sentistrength and Tempo-Sentiscore

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

3.3.3. Results

The experiment for 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 arise the difficulty in evaluating the performance. To evaluate the effectiveness of proposed system, we employed the experiments to overlap with the gold standard.

For the star rating of any entity, Sentiscore plays a vital role. From table 3, 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 ford'07 is rated as “**”. On the other hand, with temporal tagging, it is rated as “*”.

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 fig.13. Linguistic experts agreement helped in showing the Tempo-Sentiscore represent the real scenario of the opinion about any entity.

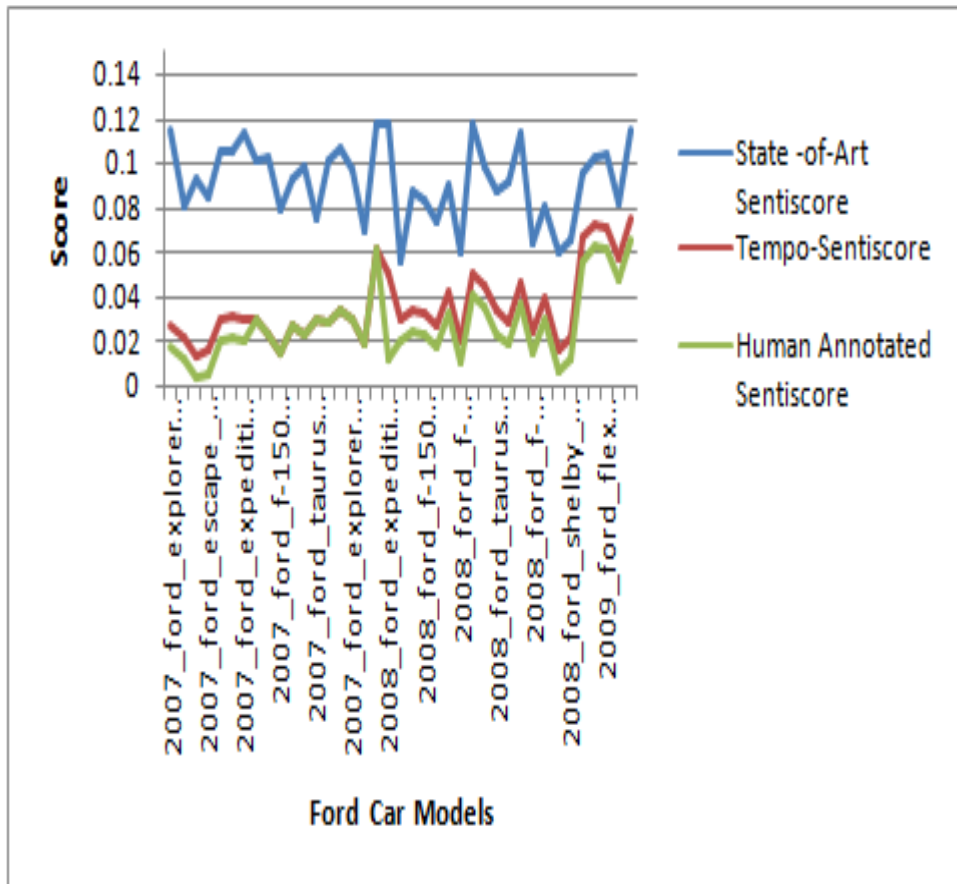


Fig. 13: Effectiveness of the Proposed System

3.3.4. Conclusion

To conclude the results, it is noticed that tempo-sentiscore outperforms the standard sentiscore. It captures the essence of temporality along with sentiment of the entity.

3.4. ANALYSIS OF VARIOUS SUPERVISED LEARNING APPROACHES

3.4.1. Introduction

There is no restriction over the use of any particular machine learning algorithm. We need to test different algorithms on the same dataset. Precision and recall values used by various algorithms gives the choice of machine learning algorithm best suited to that particular domain. In sentiment analysis, most of the work is based on paired classification of dataset i.e. +ve or -ve for sentiment analysis. Algorithms used for binary classification are Naive

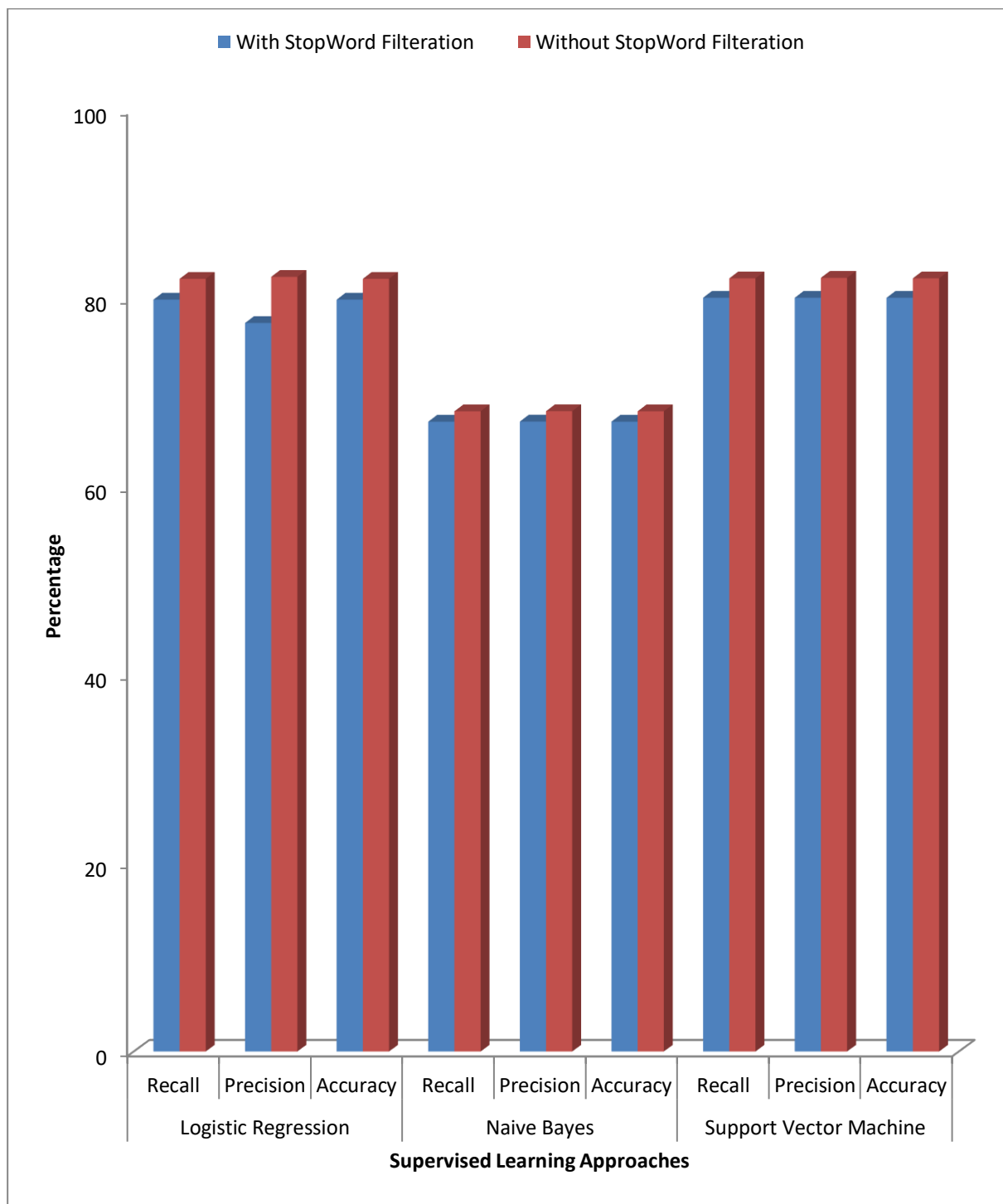


Fig. 15: Comparison of Supervised learning approaches using dataset 2

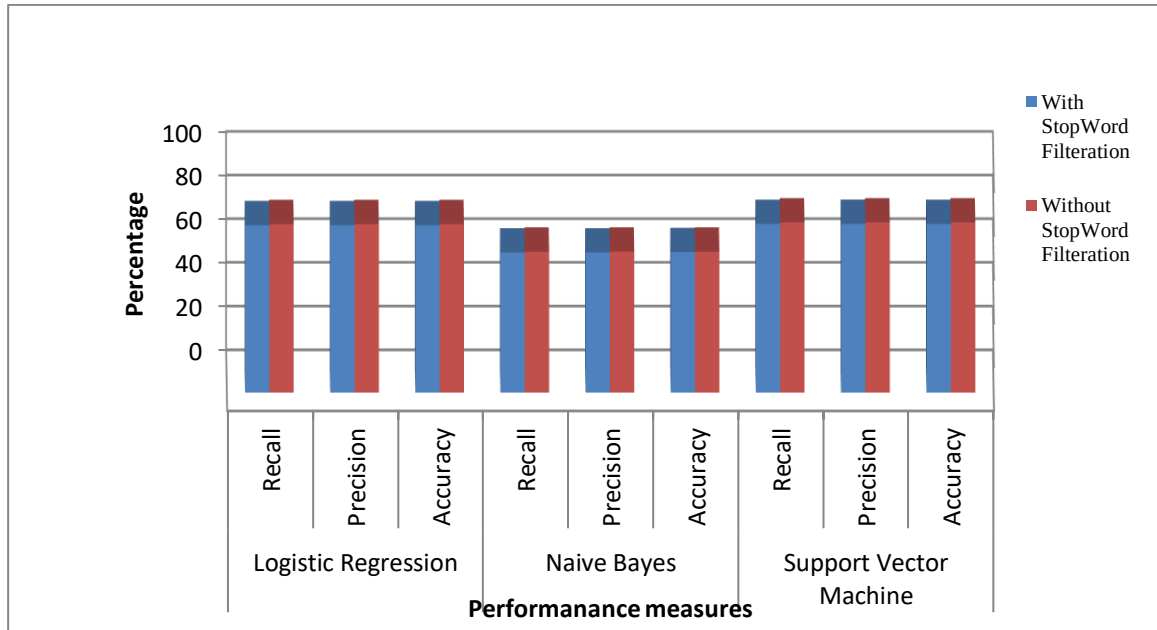


Fig. 14: Comparison of Supervised learning approaches using dataset 2

Bayes, SVM and Logistic Regression.

3.4.2. Results and discussion

Fig.14 shows the Precision, Recall and Accuracy for binary classification of movie reviews with the impact of stop words for the same dataset. For our experiment, we have used three machine learning algorithms i.e. Logistic Regression(LR), Naive Bayes(NB) and SVM . Fig.15 and Fig. 16 shows the performance of given learning algorithms for dataset 1 and dataset 2 respectively.

3.4.3. Conclusion

It is worth trying logistic regression for the movie review dataset. In the literature, researchers have used Naive Bayes, SVM, maximum entropy, etc. After the analysis of various results, we have found that logistic regression outperformed. From various aspects the results have been analysed i.e. the volume of data and stop word filtration. We have presented supervised learning scheme with 10-fold cross validation. Evaluation of results has shown that Logistic Regression is better than Naive Bayes for binary classification of the data.

3.5. CONCLUSION AND FUTURE SCOPE

3.5.1. CONCLUSION

To recapitulate, we formulate the following concluding points:

3.5.1.1 INCREASE IN RELIABILITY

We have determined that the contribution of objectives increase the reliability of the system to a great deal. It includes trimming to unwanted data or obsolete data up to an appropriate level. In some cases, less obsolete 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 tempo-sentiscore is reliable.

3.5.1.2. 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 filtration affect the results. Preprocessing also includes to deal with unstructured data contains slangs, syntactically weak structure.

3.5.1.3. DOMAIN INDEPENDENCE

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

3.5.1.4. 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.

3.5.2. FUTURE WORK

We further plan to develop a system to handle with more than two languages as a macaronic text. We are trying to map it with UNL. Along with this, we would like to elaborate the temporality by increasing the granularity of tempo-sentiscore. Along with this, We plan to work on processing of long tail words which may change the magnitude of sentiscore. This intensity.

Summary

	Work Done	State of the Art Approach	Improvement	Future Work
Objective 1	Normalization of textual data	Give equal importance to each emoticons irrespective to their meaning.	Apply Semantic analysis for mapping of each emoticons to English word.	Will apply new techniques for fake reviews identification
Objective 2	Capturing Hidden temporality from the textual data	Used Meta data i.e. date of post, for feeding into temporal values	Formulate rules using Linguistic and Metadata approach.	Increase the granularity of the processing.
Objective 3	Handling Macaronic Language Content	Neglect foreign language content from the processing.	Macaronic parser along with normalization	Processing of long tail words
Objective 4	Choice and analysis of various supervised learning approaches	-	Found that how the volume of data affect the performance.	Implement Adaboosting for model essembling.

REFERENCES

REFERENCES

- [1] Liu, Bing. "Sentiment Analysis and Subjectivity", Handbook of natural language processing ,vol.2, 2010, pp.627-666.
- [2] B. Pang, L. Lee, et al., "Opinion mining and sentiment analysis", Foundations and Trends R in Information Retrieval vol.2 no.1, 2008, pp.1-13.
- [3] Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U. and de Jong, F, "Multi-lingual support for lexicon-based sentiment analysis guided by semantics", Decision Support Systems, Vol. 62, June, 2014, pp.43–53.
- [4] Lin, Z., Tan, S., Cheng, X., Xu, X. and Shi, W. , "Effective and efficient? Bilingual sentiment lexicon extraction using collocation alignment", Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), NY, USA, 2012, pp.1542–1546.
- [5] McCord, M. and Chuah, M. , "Spam detection on twitter using traditional classifiers", Proceedings of 8th International Conference on Autonomic and Trusted Computing, Banff, Canada, 2011, pp.175–186.
- [6] Zhu, S., Liu, Y., Liu, M. and Tian, P., "Research on feature extraction from Chinese Text for opinion mining", Proceedings of International Conference on Asian Languages Processing, Singapore, 2009, pp.7–10.
- [7] Ljubešić, N., Erjavec, T. & Fišer, D., "Standardizing tweets with character-level machine translation", Computational Linguistics and Intelligent Text Processing. Springer, 2014, pp.1-12.
- [8] Yvon, F. O., "Rewriting the orthography of SMS messages", Natural Language Engineering, Vol. 16, 2010, pp. 133-159.
- [9] AW, A., Zhang, M., Xiao, J. & Su, J., "A phrase-based statistical model for SMS text normalization" , Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, 2006, pp.33-40.
- [10] Pennell, Deana L., and Yang Liu. "Evaluating the effect of normalizing informal text on TTS output." In Spoken Language Technology Workshop (SLT), 2012 IEEE, pp. 479-483.
- [11] Wang, P. & Ng, H. T. , "A Beam-Search Decoder for Normalization of Social Media Text with Application to Machine Translation", HLT-NAACL, 2013, pp.471-481.
- [12] Baldwin, Tyler, and Yunyao Li., "An in-depth analysis of the effect of text normalization in social media." , In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 420-429.
- [13] Zhang C, Baldwin T, Ho H, Kimelfeld B, Li Y, "Adaptive parser-centric text normalization", In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2013, pp. 1159-1168.
- [14] Chieu, H. L. & Ng, H. T., "A maximum entropy approach to information extraction from semi-structured and free text", AAAI/IAAI, 2002, pp.786-791
- [15] Nasukawa, T., Punjani, D., Roy, S., Subramaniam, L. V., Takeuchi, H. & Block, I. , "Adding sentence boundaries to conversational speech transcriptions using noisily labelled examples", Proc. of AND07 Workshop in conjunction with IJCAI, 2007, pp.1-8.
- [16] V. Hatzivassiloglou and K. R. McKeown, "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, USA,1997, pp. 174-181.
- [17] B. Pang, L. Lee, and S. Vaithyanathan, "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,USA, , pp. 79-86.
 - [18] B. Pang, L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in Proceedings of the 42nd annual meeting on Association for Computational Linguistics,USA, 2004,p. 271.
 - [19] B. Pang and L. Lee, "Opinion mining and sentiment analysis", Foundations and trends in information retrieval, vol. 2, pp. 1-135, 2008.
 - [20] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in LREc, Malta,2010, pp. 1320-1326.
 - [21] Alec Go, Lei Huang, Richa Bhayani, "Twitter Sentiment Analysis", Project Report, standford,2009.
 - [22] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, vol. 11,USA, 2010, pp. 122-129.
 - [23] X. Yang and W. Liang, "An n-gram-and-wikipedia joint approach to natural language identification," in Universal Communication Symposium (IUCS), 2010 4th International, China, 2010, pp. 332-339.
 - [24] S. Carter, W. Weerkamp, and M. Tsagkias, "Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text," Language Resources and Evaluation, vol. 47, 2012, pp. 195-215.
 - [25] M. Lui and T. Baldwin, "Cross-domain feature selection for language identification," in Proceedings of 5th International Joint Conference on Natural Language Processing, Thailand, 2011, pp. 553–561.
 - [26] Kaur and R. Mohana, "A roadmap of sentiment analysis and its research directions," International Journal of Knowledge and Learning, vol. 10, 2015, pp. 296-323.
 - [27] B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith, "From tweets to olls: Linking text sentiment to public opinion time series", ICWSM 11, 2010, pp.122-129
 - [28] M. Thelwall, K. Buckley, G. Paltoglou, "Sentiment strength detection for the social web", Journal of the Association for Information Science and Technology, vol.63 no.1, 2012, pp.163-173.
 - [29] Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. "Sentiment strength detection in short informal text." Journal of the American Society for Information Science and Technology, vol.62, no. 2 2011,pp. 419.
 - [30] C. Strapparava, R. Mihalcea, "Learning to identify emotions in text", in:Proceedings of the 2008 ACM symposium on Applied computing, ACM,2008, pp. 1556-1560.
 - [31] M. P. Marcus, M. A. Marcinkiewicz, B. Santorini, "Building a large an notated corpus of english: The penn Treebank", Computational linguistics vol.19 no.2,1993, pp.313-330.
 - [32] S. Baccianella, A. Esuli, F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining", in: LREC, vol. 10, 2010, pp. 2200-2204.
 - [33] B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series", ICWSM 11, 2010, pp. 122-129,
 - [34] M. Thelwall, K. Buckley, G. Paltoglou, "Sentiment in twitter events", Journal of the American Society for Information Science and Technology vol. 62 no.2, 2011, pp.406-418.
 - [35] B. Han, A. Lavie, "A framework for resolution of time in natural language",ACM

- Transactions on Asian Language Information Processing (TALIP), vol.3 no.1,2004, pp. 11-32.
- [36] X. Chang, C. D. Manning, “Sutime: Evaluation in tempeval-3”, in: SemEval@ NAACL-HLT, 2013, pp. 78-82.
 - [37] G. H. Dias, M. Hasanuzzaman, S. Ferrari, Y. Mathet, “Tempowordnet for sentence time tagging”, in: Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 833-838.
 - [38] H. Razavi, S. Matwin, J. De Koninck, R. R. Amini, “Dream sentiment analysis using second order soft co-occurrences (sosco) and time course representations”, Journal of Intelligent Information Systems vol.42 no.3, 2014, pp. 393-413.
 - [39] T. Fukuhara, H. Nakagawa, T. Nishida, “Understanding sentiment of people from news articles: Temporal sentiment analysis of social events”, in: ICWSM, 2007.
 - [40] Y. Zhu, S. Newsam, “Spatio-temporal sentiment hotspot detection using geotagged photos”, in: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2016, pp. 76.
 - [41] E. Haddi, X. Liu, Y. Shi, “The role of text pre-processing in sentiment analysis”, Procedia Computer Science ,vol.17 , 2013, pp.26-32.
 - [42] [17] B. Agarwal, N. Mittal, “Machine learning approach for sentiment analysis” in: Prominent feature extraction for sentiment analysis, Springer, 2016, pp.21-45.
 - [43] [18] M. Abdul-Mageed, M. Diab, S. Kubler, “Samar: Subjectivity and sentiment analysis for arabic social media”, Computer Speech & Language vol. 28 no.1, 2014, pp. 20-37.
 - [44] C. Cheng, J. Xu, “A sentiment analysis model based on temporal characteristics of travel blogs”, Data Analysis and Knowledge Discovery vol.1 no.2, 2017, pp.87-95.
 - [45] K. Ganesan, C. Zhai, “Opinion-based entity ranking”, Information retrieval, vol.15 no.2,2012,pp. 116-150.
 - [46] K. Almgren, J. Lee, “An empirical comparison of influence measurements for social network analysis”, Social Network Analysis and Mining vol.6 no.1,2016, pp. 52
 - [47] Agarwal, R. and Srikant, R., “Fast algorithms for mining association rules in large databases”, Proceedings of 20th International Conference on VLDB, vol. 1215, 1994, pp.487–499.
 - [48] An, N.T.T. and Hagiwara, M, “Adjective-based estimation of short sentence’s impression”, International Conference on Kansei Engineering and Emotion Research,11–13 June, Keer, Linköping, 2014, pp.1–16.
 - [49] Baccianella, S., Esuli, A. and Sebastiani, F., “ SENTIWORDNET 3.0: an enhanced lexical resource for sentiment analysis and opinion mining”, Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC’10), European Language Resources Association (ELRA), Valletta, Malta, 2010, pp.2200–2204.
 - [50] Banea, C., Mihalcea, R. and Wiebe, J. , “Multilingual sentiment and subjectivity analysis” , Multilingual Natural Language Processing, vol. 6, 2011, pp.1–19.
 - [51] Boiy, E. and Moens, M-F., “A machine learning approach to sentiment analysis in multilingual web texts”, Journal Information Retrieval, vol. 12, no. 5, October,2009, pp.526–558.
 - [52] Bollegala, D., Weir, D. and Carroll, J., “Cross-domain sentiment classification using a sentiment sensitive thesaurus”, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 8, August, 2013,pp.1719–1731.
 - [53] Brill, E. , “Some advances in transformation-based part of speech tagging”, Proceedings of the 12th National Conference on Artificial Intelligence, AAAI Press, Menlo Park, CA,1994, pp.722–727.

- [71] Kucuktunc, O., Barla Cambazoglu, B., Weber, I. and Ferhatosmanoglu, H. , “A large-scale sentiment analysis for yahoo! Answers”, Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, NY, USA,2012, pp.633–642.
- [72] Larsen, B. and Aone, C., “Fast and effective text mining using linear-time document clustering”, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA,1999, pp.16–22.
- [73] Lim, E-P., Nguyen, V-A., Jindal, N., Liu, B. and Lauw, H.W., “Detecting product review spammers using rating behaviours”, Proceedings of the 19th ACM International Conference on Information and Knowledge Management, NY, USA, 2010, pp.939–948.
- [74] Lin, Z., Tan, S., Cheng, X., Xu, X. and Shi, W. , “Effective and efficient? Bilingual sentiment lexicon extraction using collocation alignment”, Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), NY, USA, 2012, pp.1542–1546.
- [75] Liu, B. , “Sentiment Analysis and opinion Mining”, Morgan & Claypool, May 2012,p.167.
- [76] Liu, F., Weng, F. and Jiang, X. , “A broad-coverage normalization system for social media language”, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea,2012, pp.1035–1044.
- [77] Liu, L., Kang, J., Yu, J. and Wang, Z. , “A comparative study on unsupervised feature selection methods for text clustering”, Proceedings of the 2005 12th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE’05), Wuhan, China, 2005, pp.597–601.
- [78] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C. , “Learning word vectors for sentiment analysis”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2011, pp.142–150.
- [79] McCallum, A. and Nigam, K. (1998) ‘A comparison of event models for Naive Bayes text classification’, Proceedings of the AAI- Workshop on Earning for Text Categorization,vol. 752,1998, pp.41–48.
- [80] McCord, M. and Chuah, M. , “Spam detection on twitter using traditional classifiers”, Proceedings of 8th International Conference on Autonomic and Trusted Computing, Banff, Canada, 2011,pp.175–186.
- [81] Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. , “WordNet: an online lexical database”, International Journal of Lexicography, Vol. 3, No. 4, 1990, pp.235–244.
- [82] Moghaddam, S. and Popowich, F. , “Opinion Polarity Identification through Adjectives”, CoRR, abs/1011.4623,2010.
- [83] Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M. and Ghosh, R. , “Spotting opinion spammers using behavioral footprints”, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Illinois, USA,2013, pp.632–640.
- [84] Mukherjee, A., Liu, B. and Glance, N. , “Spotting fake reviewer groups in consumer reviews”, International World Wide Web Conference (WWW-2012), Lyon, France, 2012, pp.191–200.
- [85] Nitin, J. and Liu, B. , “Opinion spam and analysis”, Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, New York, USA, 2008, pp.219–230.
- [86] O’Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A., “From tweets to polls: linking text sentiment to public opinion time series”, Proceedings of the Fourth International AAI Conference on Weblogs and Social Media, Washington DC,2010,pp.1–8.
- [87] Pang, B. and Lee, L. , “A sentiment education: sentiment analysis using subjectivity summarization based on minimum cuts”, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Article No. 271, PA, USA, 2004, pp.271–278.

- [88] Pang, B., Lee, L. and Vaithyanathan, S., “Thumbs up? Sentiment classification using machine learning techniques”, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, vol. 10, 2002, pp.79–86.
- [89] Pontiki, M., Galanis, D., Pavlopoulos, I., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S., “SemEval 2014 Task 4: aspect based sentiment analysis”, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) at (COLING 2014), Dublin, Ireland, 2014, pp.27–35.
- [90] 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, 2013, pp.393–414.
- [91] Riloff, E. and Wiebe, J., “Learning extraction pattern for subjective expressions”, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03), Sapporo, Japan, 2003.
- [92] Srivastava, R., Bhatia, M.P.S., Srivastava, H.K. and Sahu, C.P. , “Exploiting grammatical dependencies for fine-grained opinion mining”, Proceedings of IEEE International Conference on Computer & Communication, September, Allahabad, Uttar Pradesh, 2010, pp.768–775.
- [93] Thelwall, M., Buckley, K. and Paltoglou, G. , “Sentiment in twitter events”, Journal of the American Society for Information Science and Technology, vol. 62, no. 2, February, 2011, pp.406–418.
- [94] Turney, P.D. , “Mining the Web for synonyms: PMI-IR versus LSA on TOEFL”, Proceedings of the 12th European Conference on Machine Learning, Springer-Verlag, Berlin, 2001, pp.491–502.
- [95] Turney, P.D. , “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”, Proceedings of the Association for Computational Linguistics (ACL), Stroudsburg, PA, USA, 2002, pp.417–424.
- [96] UzZaman, N. and Khan, M. , “T12: an advanced text input system with phonetic support for mobile devices”, 2nd International Conference on Proceedings of Mobile Technology, Applications and Systems, 2005, Guangzhou, 2005, pp.1–7.
- [97] Thelwall, Mike. "Gender bias in sentiment analysis." Online Information Review,, Vol. 42, No. 1 ,2018, pp. 45-57.
- [98] Wang, G., Xie, S., Liu, B. and Yu, P.S. , “Identify online store review spammers via social review graph”, Journal ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, September, 2012, pp.1–21,
- [99] 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, Vol. 23, May, 2015, pp.3–15.
- [100] Whittle, J., Simm, W., Ferrario, M-A., Frankova, K., Garton, L., Woodcock, A., Nasa, B., Binner, J. and Ariyatun, A. , “VoiceYourView: collecting real-time feedback on the design of public spaces”, Proceedings of the 12th ACM International Conference on Ubiquitous Computing, NY, USA, 2010, pp.41–50.
- [101] Willett, P. , “The porter stemming algorithm: then and now”, Electronic Library and Information Systems, 2006, vol. 40, no. 3, 2006, pp.219–223.
- [102] Wilson, T., Wiebe, J. and Hoffmann, P., “Recognizing contextual polarity in phrase-level sentiment analysis”, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 2005, pp.347–354.
- [103] Xie, S-x. and Wang, T. , “Construction of unsupervised sentiment classifier on idioms

- resources”, Journal of Central South University, Vol. 21, No. 4, April, 2014, pp.1376–1384.
- [104] Yi, J. and Niblack, W. , “Sentiment mining in webfountain”, Proceedings of the 21st International Conference on Data Engineering (ICDE), Tokyo, Japan, 2005,pp.1073–1083.
 - [105] Jacobo Rouces, Lars Borin, Nina Tahmasebi, Stian Rødven Eide . Defining a gold standard for a Swedish sentiment lexicon: Towards higher-yield text mining in the digital humanities, Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference Helsinki, Finland, March 7-9, vol. 2084, 2018.
 - [106] Zhang, T., Damerau, F. and Johnson, D., “Text chunking based on a generalization of winnow”, Journal of Machine Learning Research, vol. 2, March, 2002,pp.615–637.
 - [107] Zhao, W. and Zhou, Y. , “A template-based approach to extract product features and sentiment words”, Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2009), September, Dalian, China, 2009, pp.1–5.
 - [108] Zhu, J., Zhang, C. and Ma, M.Y. , “Multi-aspect rating inference with aspect-based segmentation”, IEEE Transactions on Affective Computing, Vol. 3, No. 4, 2012, pp.469–481.
 - [109] Zhu, S., Liu, Y., Liu, M. and Tian, P. , “Research on feature extraction from Chinese Text for opinion mining”, Proceedings of International Conference on Asian Languages Processing, Singapore, 2009, pp.7–10.

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, Springer, INDIA, pp.1-9, 2018. (SCI)
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. (Scopus)) (SJR-0.17)

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, Bangalore, India, July, 2016.
2. Sukhnandan Kaur, Rajni Mohana, “Prediction of Sentiment from Textual Data Using Logistic Regression Based on Stop Word Filtration 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.