# Prediction of Sentiment from Macaronic Reviews

Sukhnandan Kaur and Rajni Mohana
Department of CSE, JUIT, Waknaghat, 173234, India
E-mail: sukhnandan.kaur@mail.juit.ac.in, rajni.mohana@juit.ac.in

**Technical paper**

**Keywords:** macaronic language, sentiment analysis, supervised learning, normalization

**Received:** March 11, 2017

*Web-sphere is the vast ocean of data. It allows its users to write their opinion, suggestions over various social platforms. The users often prefer to write in their native language or some hybrid content (i.e., combination of two or more languages). It's also observed that people use a word or two of their native language in a text of base language. The presence of native words along with base language is known as macaronic languages. For example: Dunglish (Dutch and English), Chinglish (Chinese and English), Hinglish (Hindi and English) The use of macaronic languages over the web is on the rise these days. This type of text generally doesn't follow any syntactic structure, thus making processing of the content difficult. This paper deals with extracting meaningful information of a text containing macaronic content. It also facilitates the need of expert analysers for the processing of such content to take effective decisions. The performance of various decision support systems is dependable over these analysers. Therefore, this paper presents an algorithm which initially normalizes the content to its base language; later performs sentiment analysis over it. The experimental results using proposed algorithm indicates a trade-off between various performance aspects.*

*Povzetek: Prispevek predstavi iskanje razumevanja makaronskega besedila, tj. besedila z dodanimi besedami drugega jezika.*

## 1 Introduction

Online review communities successfully allow its users to write their opinion, suggestions over various social platforms. These reviews greatly affect the decision to buy or sell any product and to use any service. It is fruitful to the manufacturer or service provider to enhance the productivity. Automatic decision support systems take these reviews into account for sentiment analysis. However, it is extremely difficult to have reviews in uniform language. During an automatic processing of reviews written online, it is found that 2/3 of the internet users are non-English [5] . The reason behind this is that most of the people have the ability to learn only 2 or 3 languages proficiently. In this technological world, people have equal priority to write over the internet among different languages. People who write reviews belong to different communities from different regions of the world; they have the freedom to use their native language too. When a text contains more than one language, it is called as multilingual text. If a single sentence contains more than one language, then it is called as macaronic text[18].

Example 1: Samsung अच्छा cellphone ,

In the above mentioned text, it is taken as macaronic content containing Hindi and English languages.

These irregularities found in the data over the internet make the processing more complex. Due to the scarcity of the language resources over the web, it becomes very difficult to handle all the possible languages over the globe. It is a challenging task of a natural language processing group. The formalism in sentiment analysis limits the system to specific users. The reviews from all the users of a particular entity are valuable. It increases the need of automated systems to handle multilingual content. Derkacz et.al.[12] stated some of the requirements to have a multilingual automated system. These requirements are further taken care by language processors to build a multilingual system. In case multilingual systems, the language of whole document is taken into account whereas for macaronic language processing, we need to detect the language of each word. This paper proposes a sentiment analyser which deals with the macaronic text. Initially, reviews are to be normalized during pre-processing stage. Later, these reviews are processed through sentiment analyser.

This paper is organised as: section 2 describes the state of the art sentiment analysers. In section 3, system design and algorithm is proposed. Experimental analysis using various performance metrics are presented in section 4. Finally, the whole work is concluded in section 5.

# 2 Related work

Numerous researchers have worked in the field of natural language processing. Kaur et.al.[14] presented sentiment analysis of reviews written in Punjabi language. The researchers collected the reviews written in Punjabi which afterwards segregated into positive or negative reviews. Das et.al.[8] found the need of having SentiWordNet for Bengali language. Their work helped the researchers in the field of sentiment analysis. The researchers annotated the required lexicon. Das et.al.[7] worked for sentiment analysis of reviews written in Bengali language. In this paper, the researchers have used support vector machine (SVM) with Bengali SentiWordNet. The paper presents the feature extraction for Bengali language. Das et.al.[6] developed subjectivity clues based on theme detection techniques. Bengali corpus is used in their work and later compared the results with English subjectivity detection. Das et.al.[9] developed a gaming theory by which researchers can easily build the SentiWordNet in the required language. This work demands the respective linguistic experts. Joshi et.al.[13] used supervised learning approach for their work by using Hindi- SentiWordNet for their work. In this paper, researchers used standard translation techniques to preserve the polarity of each document while translating it. Bakliwal et.al.[2] worked for detecting subjectivity based on graph theory. Researchers explored the effect of synonym and antonym over the subjective nature of the document. The results were good for Hindi and English. The researchers claimed that their strategy will work well in other languages too. Das et.al.[10] developed a system for deducing the emotion and intensity of emotion based on sentiment hidden in the data. In this work, researchers have used supervised learning methods for their work. Richa et.al.[21] presented a survey for sentiment analysis in Hindi language. The results have shown that sentiment analysis in Hindi language is complex as compared to English language. The reason behind this complexity is the non-uniform nature of the Hindi language. Various research challenges are also discussed. Researchers[21] developed a system which depicts the polarity of the text and tested their system over the Hindi movie reviews. Parul et.al.[1] developed a sentiment analyser for movie reviews written in Punjabi language using various machine learning algorithms. Raksha et.al.[20] used semi-supervised technique for polarity detection in Hindi movie reviews. In their work, researchers reported $87\%$ accuracy of the proposed system by using bootstrapping and graph based approach for sentiment analysis. Pooja et.al.[17] used Hindi SentiWordNet for finding opinion orientation of the reviews. Researchers used unsupervised learning for their work. Kerstin et.al.[11] developed a system for multilingual text for obtaining the polarity of reviews written in language other than resource rich language English. Researchers used a standard translation methodology and supervised learning for sentiment analysis. C. Banea et.al.[3] developed a system which focused on the sentiment analysis based on translation of input document other than English. In their work, researchers used English as a source language. They used supervised learning approach for their work. For the translation of the text correctly various available translators are used. i.e. Goggle, Moses, Bing translators.

The work by different researchers is summarized into table 1 . It is noticed that researchers are focusing well in the area of multilingual sentiment analysis. Researchers focused in finding document language for translating any document into base language instead of language of individual word. This sometimes discard the opinion bearing word written in any foreign language. As in example 1, the word अच्छा, means good is discarded if the document language is detected as English. The efficient processing of such documents is required to increase the effectiveness of the decision support system.

## 2.1 Motivation

After looking into the scenario, we found that we need SentiWordNet in almost every language all over the global. It is very complex task. The motivation behind the proposed system is that the existing system for multilingual sentiment analysis is unable to process macaronic data. The rise in the volume of macaronic data over the internet arise the need of proposed system. The reasons for having macaronic content over the web in huge volume are as follows:

1. Scarcity of Resources: Sentiment analysis task demands for the availability of lexicons or data of any particular language. There is huge variation in every language model. This makes the model used for one language cannot be used for other languages.
   For example: Chinese language model does not consider spaces while as other models focus mainly over spaces to tokenize.

2. Lack of uniformity of languages: Most of the languages often follow their own traditional structures. Thus, processing of each language data with the general structure model gives unsatisfactory results.
   For example: English language use Subject-Verb-Object(SVO) while Hindi Language model follow Subject-Object-Verb (SOV)

3. Freedom of writing in native language: People these days have number of followers from different countries through various online applications. They are also able to propagate their ideas through it. Sometimes, few words they prefer writing in their own native language, which may not be understandable by some of the followers. In case of an automated system, during pre-processing through one language model, these native words may be neglected taken as foreign language words. Sometimes, we may lose meaningful information during this type of pre-processing.
   For example: सैमसंग is on great demand.
   सैमसंग(Samsung) is negected by English language

| Author | Work | Level | Language | Results | Technique | Corpus | Year |
|---|---|---|---|---|---|---|---|
| Danet et.al.[5] | Classification of reviews into positive or negative opinion | Document level | Punjabi | Accuracy = 75% | Machine Learning | Blogs | 2014 |
| Derkacz et.al.[18] | 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.[14] | 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 |
| Bandyopadhyay et. al.[6] | 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.[9] | Subjectivity clues based on antonym and synonym using graph theory | Document Level | Hindi, English | Accuracy = 79% | Supervised | Movie reviews | 2012 |
| Sharma et. al.[10] | Polarity detection of movie reviews using unsupervised techniques | Sentence Level | Punjabi | NA | Unsupervised | Movie reviews | 2015 |
| Arora et. al.[21] | Sentiment orientation of reviews written in Hindi language | Document Level | Hindi | Precision = 70.04%, Recall = 63.02% | Unsupervised | Movie reviews | 2014 |
| Sharma et. al.[1] | Sentiment analysis using Semi-Supervised techniques | Document Level | Hindi | Accuracy = 87% | Semi-Supervised | Movie reviews | 2014 |
| Pandey et. al.[20] | Opinion orientation of Hindi movie reviews is deduced using Hindi-WordNet | Document Level | Hindi | NA | Unsupervised | Movie reviews | 2015 |
| Denecke et. al.[17] | Polarity detection from reviews using standard translation of German reviews in English afterwards find the polarity | Document Level | German, English | Accuracy = 66% | Supervised | Movie review | 2008 |
| Banea et. al.[11] | Enabling Multilingual question answering system | Document Level | French, German and Spanish | NA | Supervised | Question Answers | 2016 |

Table 1: State of Art Multilingual Sentiment Analysis

model. Thus, it becomes difficult to extract samsung as an entity.

4. For getting point of attraction: People use the multilingual content or some fancy words in various applications like product advertisements, shop names, etc. This makes the task of processing such web content complex. For example:
सamसung (Samsung) is on great demand. ■
सamsung (Samsung) is on great demand.
म ona(Mona) is feeling so good.
Hence, from the above examples, Samsung is hard to detect as it is being neglected by chosen language model.

Due to the above mentioned reasons, it is very much necessary to have an efficient system to process macaronic language content. Our contribution is to enhance the performance i.e.precision, recall and accuracy using supervised sentiment analysers. The proposed system is with less fallout which shows its high efficiency.

# 3 System design

The proposed system as shown in Figure 1 applies a variant of techniques for normalization of macaronic text and classification of reviews. The system consists three major components:

1. Language Processing
2. Text Processing and
3. Sentiment Analysis

A component based on language detection is carried out using algorithm 1. The core idea of this component is to normalize the macaronic content. Other two components are carried out using algorithm 2. It normalizes the content to extract the SentiStrength of each document. Combination of these two algorithms (Algorithm 1 and Algorithm 2) is used to carry out sentiment analysis for multilingual or macaronic language documents.

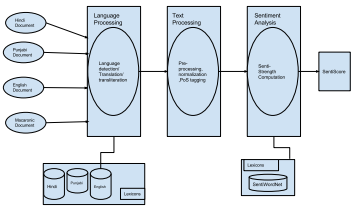



Figure 1: Proposed System Design

1. Language Processing: It is the primary component of the proposed system. In this component tokenization, language detection and conversion of tokens to its base language is carried out. These sub-components are described as follows:

   (a) Tokenization: It is the basic unit of any language processing task. A sequence of sentences, words or characters are passed as an input to any system.
   The output of this phase is tokens. It can be done at different levels depending upon the level of granularity: sentence level, word level, character level as shown in table 2. The proposed system is based on word level tokenization for macaronic language.
   E.g. Samsung has a good market value. Users are happy with its mobile products.

   (b) Language Detection/ Translation: For language detection, we have used PoS[19] tagging, as shown in table 3. The unrecognized or untagged tokens can be passed through language detection module. The output of this phase is the tokens in the base language of the system. i.e. Taking English as a base language. If the token is found in Hindi WordNet then Hindi to English translator is applied to it. On the other end, if the word belongs to Punjabi language, it is passed through the Punjabi to English translator. It is a general procedure which can be applied to various other languages too.

2. Text Processing: It is the second important component of the proposed system. It carries various sub-tasks described as follows:

   (a) Normalization: After filtration of subjective sentences, normalization is to be done. The process of normalization is to regularize or process the grammatical variants present in the sentence. Grammatical variants include past verbs (regular and irregular) / present verbs, classification of noun phrases in singular and plural. In normalization, finding the abbreviations, case folding, etc.Normalization is a process having data in a well format as required for appropriate processing. It includes:

| Level of Processing | Number of Tokens |
|---|---|
| Sentence Level | 2 |
| Word Level | 13 |
| Character Level | 74 |

Table 2: Tokenization at different levels

i. Handling Slangs: Slangs are playing indispensable role in opinion mining. So, it will be worthless to reject all the slangs by counting them as stop words. Various algorithms are applied to handle different types of slangs.Types of slangs[5]:
   - Emoticons: Bad☹, happy☺
   - Interjections:Mmmmm-pleasure, hmmmm-wondering, Mhmmm-confirmation
   - Intensionally misspelled: cooooooool, goooooooood, nyt, etc
   - Alphanumeric strings: gr8, 9t, etc.

   Test sentence:
   She is flying high by having this cellphone.☺
   She is flying high by having this cellphone. Happy

ii. Idiomization / Replacement of idioms with their actual meaning: In English literature, idioms play very important role in fixing the opinion from sentence about the particular entity. If the stops words are removed then some words which may or may not the part of the idiom can be rejected. In reality, these words are highly contributed to the opinion.
   Test sentence:
   She is flying high by having this cellphone. Happy
   She is very happy by having this cellphone. Happy

(b) Tokenization: In our work, we have used word level tokenizer as mentioned in table 2 . The reason behind this to process each token according to its own language instead according of language of the document.

(c) PoS Tagging: Part of speech tagging plays a vital role in natural Language processing tasks. Initially, we have tried to focus whether the state of art PoS taggers are able to recognise a foreign word. For this purpose, we have used NLTK tagger[15] and Stanford Tagger[16]. We have shown the results of both the taggers for various test sentences in table3. We have found various untagged tokens which are then processed through language processing phase.

3. Sentiment Analysis: In this module, the potency of each review is calculated. The magnitude of the sentiment associated with each document is calculated by aggregating all the review's sentiscore corresponding to that document. SentiWordNet is the base for getting the actual magnitude of the sentiment of a document. For our work, we have used SentiWordNet v3.0.0. Sentiscore corresponds to each document is taken as an output as shown in table4 .

■

# 4 Evaluation

## 4.1 Dataset

We have extracted a corpus of reviews of 10 movies containing 200 movie reviews i.e.100 positive and 100 negative; 160 reviews were used for training and 40 for testing. Each review has a size ranges from 500 to 1000 words. Initially, classification of the corpus is elaborated according to user′s scoring: reviews are marked between 3 and 5 star rating are classified as positive whereas reviews marked between 0 and 2 are taken as negative. This prior classification is based on the assumption that the star rating is correlated to the sentiment of the review. For experiment evaluation, the data was pre-processed with the TreeTagger5, POS tagger and lemmatization tool. We have used Support Vector Machine (SVM), Nave Bayes, kNN and convolutional network as classification models to train the system and classify movie reviews. The reviews are not monolingual. These reviews are macaronic in nature i.e. it consists of more than one language i.e. Hindi and English in a single review. We manually annotate the reviews based on language of each token. The guidelines for annotation are stipulated the need of retaining the semantic structure of tokens. Five different graduate students participated in the reviewing process to formulate Gold Standard. To evaluate the inter-personnel disagreement, we have used kappa measure[4] and score 0.61 is obtained.

## 4.2 Performance

Formally, the performance of proposed sentiment analyser, PSA is a function of four factors as follows:
PSA(l,$L_d$,t,$E_s$)
Where $L_d$ is Language Detection
l is a Learning Algorithm
t is a Tagger
$E_s$ is a Experimental Setup

The performance of the analyser is directly affected by the choice of optimal parameters for each factors mentioned above. In the case of optimal parameters choice for each of the factor, sentiment analyser gives maximum performance (PSAmax).

On the other end, training consists machine translated data and testing of the learning algorithm is based on the human annotated dataset i.e. Gold Standard. The performance of sentiment analyser (PSA) is negatively affected by error in language detection phase ($E_{Ld}$) as given in equation 1 .

$$PSA = PSA_{max} - E_{Ld} \quad (1)$$

In case of optimal parameters, $E_{Ld} \rightarrow 0$, PSA = $PSA\_max$

| 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 |

Table 3: Tagging of various test sentences using NLTK and Stanford Tagger

| Test Sentence | SentiStrength |
|---|---|
| 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: Sentiscore Associated With Review

| Metric | Target | Target |
|---|---|---|
| Selected | tp | fn |
| Selected | fp | tn |

Table 5: Confusion metric used to evaluate performance

## 4.3 Performance metric

For the analysis of results, the following performance metrics are used by various natural languages processing task including sentiment analysis. It includes precision, recall, F-measure and accuracy. These measures can be calculated using the confusion metric given in table 5.

Precision: It is defined as fraction of retrieved documents that are relevant. It is calculated using equation 2.

$$P = \frac{\text{number of correct positive or negative documents detected by the system}}{\text{no. of positive/negative documents detected by the system}} \quad (2)$$

Recall: It is defined as fraction of relevant documents that are retrieved. It is calculated using equation 3.

$$R = \frac{\text{number of positive or negative documents detected by the system}}{\text{no. of positive/negative documents present in the Gold Standard test set}} \quad (3)$$

F-measure: It is a harmonic mean with takes precision and recall both into account. It is a consecutive average of precision and recall. F-measure with $\alpha = 0.5$, means taking precision and recall at equal weightage.It is calculated using equation 4.

$$F = \frac{(\alpha^2 + 1) \times P \times R}{\alpha^2(P + R)} \quad (4)$$

Accuracy: it is the fraction of classifications that is correct. . It is calculated using equation 5.

$$A = \frac{t_p + t_n}{t_p + t_n + f_n + f_p} \quad (5)$$

Fall-out: It is a measure of the proportion of mistakenly selected non- targeted items. . It is calculated using equation 6

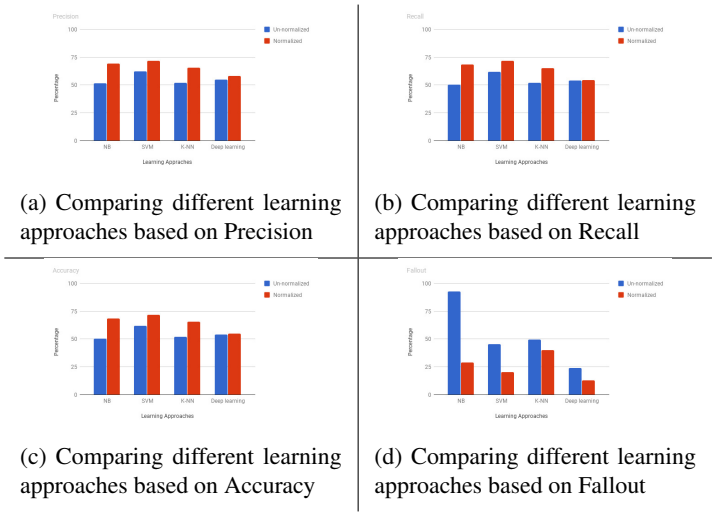
$$FO = \frac{f_p}{t_n + f_p} \quad (6)$$

## 4.4 Results and analysis

The outcomes of our experimental study are presented in Table 6 and Table 7. We can easily notice that every machine learning approach has its own pros and cons. Each of them is valuable in different aspects i.e. precision, recall, accuracy, fallout and execution time. To validate our results we have used 10-fold cross validation. For the experimental setup, we have used Support Vector Machines

| 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 |

Table 6: Un-normalized Macaronic Sentiment Analysis




(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 2: comparision of various methods

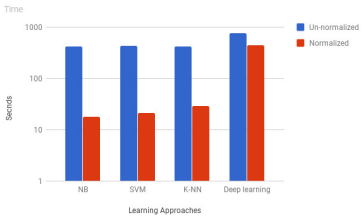



Figure 3: Comparison of execution time various machine learning Algorithms based on Proposed Scheme for normalized and unnormalized data

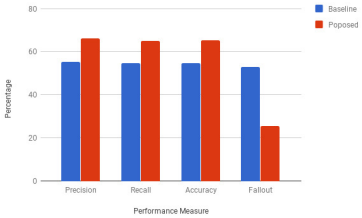



Figure 4: Comparison of proposed technique with State of art

(SVM), Nave Bayes (NB), kNN and Convolutional network(Deep Learning) to analyse the performance of proposed algorithm. The results are shown in Table 6 and Table 7. Precision, recall, accuracy, fallout are taken in percentage and time is taken in seconds. Time taken by each of the learning technique is very dependent on data size, data types, number of columns, computer hardware, memory, background running processes, cores, etc. This may vary with the change in any of the mentioned attribute. Hence, the time taken in table 6 and table 7 helped in deducing the time trend of each learning model. It is shown as an increasing order and noticed the reduction in the time to the marginal level in normalized content.

Order for unnormalized content:
$kNN < NaiveBayes < SVM < Convolutionalnetwork$
Order for normalized content:
$NaiveBayes < SVM < kNN < Convolutionalnetwork$

The results have shown in Figure 2 clearly evident the performance of proposed system using various learning approaches. These figures highlight the proposed system performance in various aspects. The proposed scheme outperforms the existing system using Nave Bayes by the rise in the values of precision, recall by 17.88% and 18.22%. Observing the results of other classifiers i.e. SVM, kNN and convolutional network also shows significant impro-

| 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 |

Table 7: Proposed normalized Macaronic Sentiment Analysis

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

Table 8: Comparison with Existing Sentiment Analysis

vements in performance levels. Using SVM and kNN more than 9% and 13% improvement is noticed in precision and recall values using proposed approach. It is also noticeable that there is a trade-off between various performance aspects. The effectiveness of system is shown by convolutional network but it takes more time than other classifiers for macaronic sentiment analysis.

Through observing Figure 3, we have found that the proposed algorithm also greatly affect the time taken by each model. It is noticeable that the normalized content reduces the training time in every learning approach. By observing Table 8, results are compared to the baseline approaches; the average value of precision, recall is increased while the fallout is decreased significantly. Figure 4 shows that how effective the proposed approach is as compared to the state of the art sentiment analysis for macaronic language.

# 5 Conclusion

Over the web where huge user generated content has already existed; the need for sensible computation for decision support system is rising. The multilingual online content has led to the increase of web debris, which is inevitably and negatively affecting information retrieval and extraction for decision support systems. To analyse this negative trend and propose possible solution, this paper focused on the evolution of sentiment analysis based on bag-of-words for macaronic reviews. Different supervised machine learning approaches gave different cross validated results. This is done by borrowing the concept of training and testing from the field of machine learning. After successful evaluation, it is concluded that there is a trade-off between various performance measures. In this study, we have investigated the need to normalize the macaronic text. We have also performed sentiment analysis over the macaronic language text consists English and Hindi. We have found an average of about 11% rise in precision and recall values. It is also noticeable that training time is also reduced significantly using proposed approach. We further plan to develop a system to handle with more than two languages as a macaronic text for sentiment analysis. We also plan to apply our proposed algorithm for entity extraction.

# References

[1] Arora, P. and B. Kaur (2015). "Sentiment Analysis of Political Reviews in Punjabi Language." International Journal of Computer Applications 126(14).

[2] Bakliwal, A., P. Arora, et al. (2012). Hindi subjective lexicon: A lexical resource for hindi polarity classification. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC).

[3] Banea, C., R. Mihalcea, et al. (2008). Multilingual subjectivity analysis using machine translation. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.

[4] Bunt, H., V. Petukhova, et al. (2016). Dialogue Act Annotation with the ISO 24617-2 Standard. Multimodal Interaction with W3C Standards, Springer: 109-135.

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

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

[7] Das, A. and S. Bandyopadhyay (2010). Opinion-Polarity Identification in Bengali. International Con-

ference on Computer Processing of Oriental Languages.

[8] Das, A. and S. Bandyopadhyay (2010). "SentiWordNet for Bangla." Knowledge Sharing Event-4: Task 2.

[9] Das, A. and S. Bandyopadhyay (2010). "SentiWordNet for Indian languages." Asian Federation for Natural Language Processing, China: 56-63.

[10] Das, D. and S. Bandyopadhyay (2010). Labeling emotion in Bengali blog corpusa fine grained tagging at sentence level. Proceedings of the 8th Workshop on Asian Language Resources.

[11] Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on, IEEE.

[12] Derkacz, J., M. a. Leszczuk, et al. Definition of Requirements for Accessing Multilingual Information and Opinions. Multimedia and Network Information Systems, Springer: 273-282.

[13] Joshi, A., A. Balamurali, et al. (2010). "A fall-back strategy for sentiment analysis in hindi: a case study." Proceedings of the 8th ICON.

[14] Kaur, A. and V. Gupta (2014). "Proposed Algorithm of Sentiment Analysis for Punjabi Text." Journal of Emerging Technologies in Web Intelligence 6(2): 180-183.

[15] Kothapalli, M., E. Sharifahmadian, et al. "Data Mining of Social Media for Analysis of Product Review." International Journal of Computer Applications 156(12).

[16] Nguyen, D. Q., D. Q. Nguyen, et al. "A robust transformation-based learning approach using ripple down rules for part-of-speech tagging." AI Communications 29(3): 409-422.

[17] Pandey, P. and S. Govilkar (2015). "A Framework for Sentiment Analysis in Hindi using HSWN." International Journal of Computer Applications 119(19).

[18] Renduchintala, A., R. Knowles, et al. "Creating interactive macaronic interfaces for language learning." ACL 2016: 133.

[19] Seih, Y.-T., S. Beier, et al. "Development and Examination of the Linguistic Category Model in a Computerized Text Analysis Method." Journal of Language and Social Psychology: 0261927X16657855.

[20] Sharma, R. and P. Bhattacharyya "A Sentiment Analyzer for Hindi Using Hindi Senti Lexicon."

[21] Sharma, R., S. Nigam, et al. (2014). "Polarity detection movie reviews in hindi language." arXiv preprint arXiv:1409.3942.

**Algorithm 1:**

**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:**
$W_s(weightedSentiStrengthofeachdocuemnt)$
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 SentiStrength
**end for**

**Algorithm 2:**

**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

**Output:**
$W_s(weightedSentiStrengthofeachdocuemnt)$
{Token list(TL) = (t1,t2,.....,tn)}
{Word List(WL)= (w1,w2,w3,.....wx)}
{'q' is the total number of tokens in a document}
{P = list of 'positive category words}
{N = list of 'negative category words}
{Pw = weight assigned to a token belongs to positive category as per SentiWordnet}
{Nw = weight assigned to a token belongs to negative category as per SentiWordnet}
Begin
**for** $d = 1$ to $k$ **do**
Tokenization
Stemming
Normalization
**for** $k = 1$ to $m$ **do**
**if** $(t_k \in W) \bigcap (t_k \in P)$ **then**
$w_{pos}(k) = Pw(t_k)$
**else if** $(t_k \in W) \bigcap (t_k \in N)$ **then**
$w_{neg}(k) = Nw(t_k)$
**else if** $(t_k \in W) \bigcap (t_k \notin N) \bigcap (t_k \notin N)$ **then**
$w_{neu}(k) = 0$
**end if**
**end for**

$$W_s = \sum_{j=1}^{m} w_{pos}(j) \pm \sum_{j=1}^{m} w_{neg}(j) \qquad (7)$$

**end for**