Unsupervised Document-Level Sentiment Analysis of Reviews Using Macaronic Parser

Sukhnandan Kaur and Rajni Mohana

Abstract Exponential rise in the multilingual web content affects the present-day decision support system to a great extent. To normalize such web content is the need of an hour. Reliability of decision support system broadly depends on the flawless processing of the language data present over the web. Macaronic text is one of the text usually found over the web. It is basically the text that contains number of languages in a single document instead of uniform language for whole document. To cope up with such a text, in this paper we propose a macaronic parser. This parser is language-independent and task-independent. The output of the proposed system is the normalized uniform base language text. This output can further be used in many other language processing tasks.

Keywords Sentiment analysis · Opinion mining · Macaronic language Multilinguality · Sentiscore · Sentistrength · Punjabi text · Hindi text

1 Introduction

Sentiment analysers 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 analysers a bit complex. Nowadays, there are large number of sentiment analysers available for different languages. To handle

S. Kaur (✉) · R. Mohana
Computer Science and Engineering, Jaypee University of Information Technology, Solan, India
e-mail: sukhnandan.kaur@mail.juit.ac.in

R. Mohana
e-mail: rajni.mohana@juit.ac.in

N.R. Shetty et al. (eds.), *Emerging Research in Computing, Information, Communication and Applications*, https://doi.org/10.1007/978-981-10-4741-1_36

multilingual text is a big challenge in sentiment analysis. 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. There have been a lot of studies on the information set that can be extracted from tweets and Facebook messages or posts. 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. We 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 recognization, pronoun resolution, automatic summarization along with sentiment analysis.

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

Example1:

This is a ਚੰਗੀ *movie.*

Here the Punjabi language word (ਚੰਗੀ) of the context is taken as garbage and tossed aside and the English portion of it will be taken into consideration. With this we lost meaningful information. Here, the opining is 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 convert everything from Punjabi 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 analysers give the neutral opinion about the movie although it is positive.

This arose the need to normalize the macaronic text. This paper proposes a method to fragment the text and autodetect 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 the particular foreign text into a base language. For our convenience, we have taken the English as a base language.

The remainder of the paper is organized as under: Section 2 describes the state-of-the-art sentiment analysis. Section 3 contains the detailed description of the proposed system. In Sect. 4, the proposed algorithm is presented. Section 5 investigates the results. Section 6, concludes the results.

2 Related Study

Various researchers are working in the field of sentiment analysis. They get huge success in their work. The task of sentiment analysis started decades ago. With time, it becomes the prime task for various enterprises to enhance their reputation in the market. Earliest works in the area of sentiment analysis is done by Hatzivassiloglou et al. [1]. They have used adjectives for deducing the polarity of the document. His work is then elaborated by Pang et al. [2]. 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 [3, 4], where they have used minicut algorithm for opinion summarization and also presented variousopinion mining techniques. Latest research paying more attention to the sentiment analysis over the data collected from various social sites. Researchers [5, 6] have used social media data for the sentiment analysis. Connor et al. [7] have used twitter data for sentiment analysis based on time series. Yang and Liang [8] (2010) proposed a new approach for identification of natural language, i.e. joint approach based on N-Gram frequency statistics and Wikipedia. Carter et al. [9] have used N-gram approach to identify five languages from the microblog. Later Lui et al. [10] followed same approach over the long and short textual documents. A detailed study of sentiment analysis at various levels of granularity [11] 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.

3 Proposed System

We have taken the following approach to isolate and identify the language before embarking up on the journey to neutralize it through translation. The system design for the macaronic paerser is given in Fig. 1. Processing of the text passes through different phases of the system.

Phase1: 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

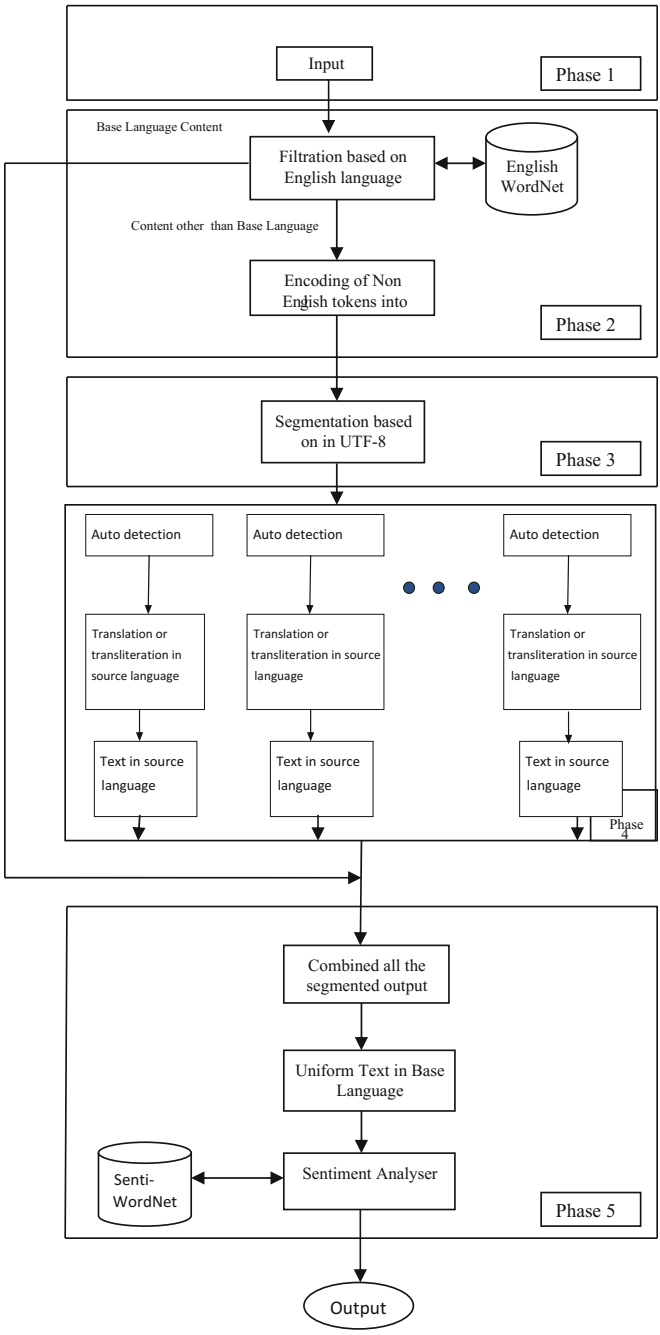

Fig. 1 Proposed system design of macaronic parser

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.

Phase3: In this phase, encoding 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. Under this, based on 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 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.

4 Proposed Algorithm

Input : Documents D = {d1, ..., dn}

'n' is the total no. of documents

Output : Document in a uniformed text ready for further language processing task.

Algorithm:

1. Set i to 1
2. for each document (d_i) in D
3. Encoding based on UTF-8 / UTF-16
4. Segmentation based on encoded document. //similar category segments are combined
5. for each segmented text
6. Detection of the language. // Hindi/English/Punjabi based on UTF-8 / UTF-16
7. Apply translation techniques
8. Content of each segment in source language.
9. End for
10. Combine all the fragmented text
11. End for
12 Apply POS tagging.
13. Processing of the formatted text for NER/ Opinion Mining/ Summarization/etc.

The proposed algorithm is very useful in dealing with the informal multilingual content present over the web. Following text describe the broad processing steps of the system.
Step 1: Extracting the stop words from the mainstream sentence. We do this by selecting a base language and running the words through a dictionary of that particular language. The index value for each token is retained. These tokens become handy while plugging back at their original place.

Input: User defined text

Output: Two list containing base language words and foreign words.

Step 2: The foreign language words are entirely different from that of English (base language in our case), and thus we understand that it is not completely useless but may or may not be informative to us. Hence, we decode the same and run it through a UTF16 to UTF 8 encoder which further gives us results for the information of the script type.

Our studies have shown that Devanagiri script when converted to UTF 8 bytes, gives us a combination of three bytes. the script range starts from [224] [164] [191] to [224] [165] [129].

Input: List containing foreign words

Output: Unicode information of the script in which the foreign words are written.

Step 3: We use this information to extract the script in which the particular text is written. We don't need any sort of prior knowledge for any language for the same. All we need to know is a Unicode set for a script and run the text by it. Using that, we will autodetect the language on its own, without the help of any language experts.

Input: Unicode information of the script in which the foreign words are written

Output: The script, in which the words are written.

Step 4: After we have extracted the information, we utilize some back ground information to finalize the language for a script. However, in this paper we limit our approach to only two languages Hindi and Punjabi language script.

Hence, we run out text to the translator API, which translates our language to the particular base language, i.e. English

Input: Foreign text along with the script information

Output: Translated or neutralized text in English (Base language)

Step 5: Plug the neutralized tokens back to the original text, to make the entire picture clear.

Input: or neutralized text in English (Base language)

Output: Full input text in one base language i.e. English in this paper.

Step 6: Part of speech tagging is done to perform the various language processing subtasks.

Input: Full input text in one base language i.e. English in this paper.

Output: POS tagged sentence.

Step 7: Depends on subtasks to be performed, the tagged text is passed to the system. As in this manuscript we have used sentiment analysis. Therefore, we have passed the text to sentiment analyser.

Input: POS tagged sentence

Output: Sentiscore associated to each document.

5 Evaluation of the System

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. In this manuscript, we have used Punjabi, Hindi words in a particular sentence. We have applied NLTK pos tagger to find the opinionated words. From Table 1, we can see that how the presence of foreign words actually affects most of the language processing tasks. In this manuscript we have analysed the system for sentiment analysis.

Table 1 Results of sentiment analyser

Test sentence (Macaronic)	Test sentence (English statements)	POS. Macaronic	POS. English	SentiStrength (Macaronic)	SentiStrength (English)
This phone has very ਵਧੀਆ battery backup	This phone has very good battery backup	This|DT phone|NN has|VBZ very|RB ਵਧੀਆ|: battery|NN backup|NN|	This|DT phone|NN has|VBZ very|RB good|JJ battery|NN	0.108	0.293
This phone has very अच्छा battery backup	This phone has very good battery backup	This|DT phone|NN has|VBZ very|RB अच्छा|: battery|NN backup|NN|	This|DT phone|NN has|VBZ very|RB good|JJ battery|NN backup|NN|	0.108	0.293
This ਫੋਨ has better battery backup	This phone has better battery backup	This|DT ਫੋਨ|NN has|VBZ better|JJR battery|NN backup|NN	This|DT phone|NN has|VBZ better|JJR battery|NN backup|NN	0.415	0.415
This दलिचस्प movie i will surely watch	This interesting movie i will surely watch	This|DT दलिचस्प|NN movie|NN i|PRP will|MD surely|RB watch|VB	This|DT interesting|JJ movie|NN i|PRP will|MD surely|RB watch|VB	0.053	0.117
ਇਸ restaurant ਦਾ ਖਾਣਾ ਬਹੁਤ good ਹੈ	The food of this restaurant is very good	ਇਸ|NN restaurant|VBD ਦਾ|CD ਖਾਣਾ|CD ਬਹੁਤ|CD good|JJ ਹੈ|NN	The|DT food|NN of|IN this|DT restaurant|NN is|VBZ very|RB good|JJ|	0.478	0.395
इस restaurant का खाना बहुत अच्छा है	The food of this restaurant is very good	इस|NN restaurant|VBD का|CD खाना|CD बहुत|CD अच्छा|CD है|CD	The|DT food|NN of|IN this|DT restaurant|NN is|VBZ very|RB good|JJ|	0	0.395

In Table 1, for the convenience top six sentences taken to show the results, SentiScore associated with the compound sentence i.e. contain foreign words from the Hindi or Punjabi language i.e. “Test Sentence (Macaronic Statements)” under the column name “SentiStrength(Macaronic)”. Other two columns show the corresponding English text and its sentiscore under the column heading “Test Sentence (English Statements)” and “SentiStrength (English)” respectively. From column “POS.Macaronic”, it can be seen that how the foreign language words in a single sentence are ignored by the present day pos taggers. We have tried other taggers also, they also gave unsatisfactory results. Most of them took the foreign language words as a noun irrespective to language. This harms the language processing results to a great extent.

We have used gold standard dataset to verify the results. In Figs. 2 and 3, it can be seen that for the reviews having same semantic structure irrespective to the language have different graph slopes. The results did not follow the actual trend set through the gold standard as shown in Fig. 3. After processing the reviews having

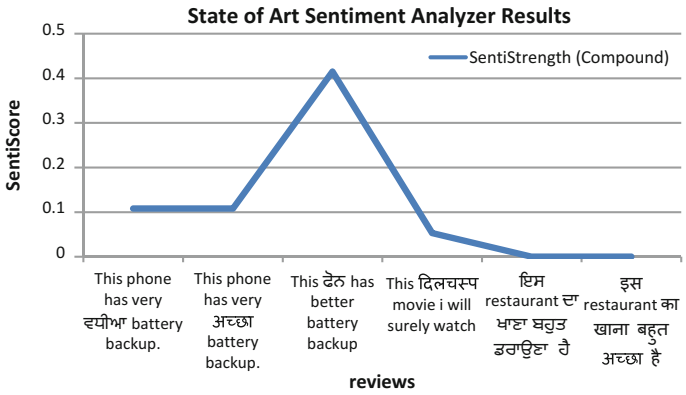

Fig. 2 Sentiscore calculated by state-of-the-art sentiment analysers

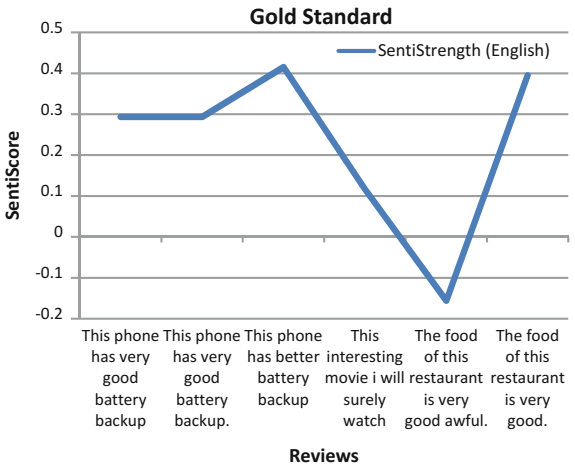

Fig. 3 Gold standard of Sentiscore

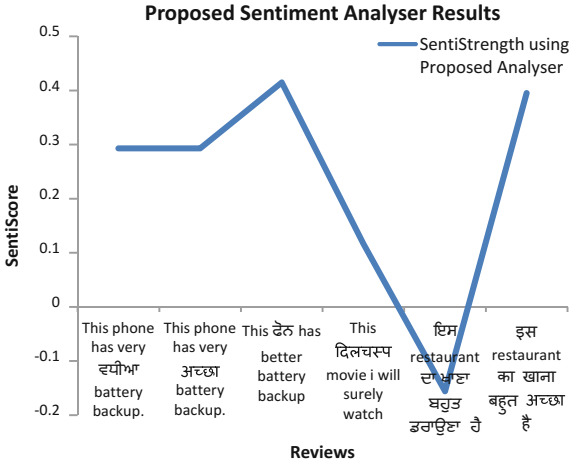

Fig. 4 Sentiscore calculated by proposed sentiment analysers

Macaronic language through the proposed sentiment analyser the results are fairly significant as shown in Fig. 4. It follows the same trend as shown in Fig. 3. This shows the significant results of the proposed system.

6 Conclusion

To recapitulate, this paper proposed macaronic parser to deal with the unnormalized web content. The results have shown the reliability of the proposed system is more as compared to the existing system. The sentiscore generated through proposed system is much closer to the gold standard of that particular dataset. Results suggest that this type of text processing is highly beneficial for the Named Entity Extraction, Feature Extraction, E-mail summarization, etc. To summarize the results, we can say that there is still a lot to do for resource scarce Indian languages to increase the feasibility of the decision support system.

References

1. 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*, 1997, pp. 174–181.
2. 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, pp. 79–86.

3. B. Pang and 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*, 2004, p. 271.
4. B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval,* vol. 2, pp. 1–135, 2008.
5. A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *LREc*, pp. 1320–1326.
6. A. Go, L. Huang, and R. Bhayani, "Twitter sentiment analysis," *Entropy,* vol. 17, 2009.
7. B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," *ICWSM,* vol. 11, p. 1.2.
8. X. Yang and W. Liang, "An n-gram-and-wikipedia joint approach to natural language identification," in *Universal Communication Symposium (IUCS), 2010 4th International*, pp. 332–339.
9. 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, pp. 195–215.
10. M. Lui and T. Baldwin, "Cross-domain feature selection for language identification," in *In Proceedings of 5th International Joint Conference on Natural Language Processing*.
11. S. Kaur and R. Mohana, "A roadmap of sentiment analysis and its research directions," *International Journal of Knowledge and Learning,* vol. 10, pp. 296–323.