Analysis on Code-Mixed Data for Movie Reviews

Sannidhi Shetty[1], Shruti Nair[2], Shruti Khairnar[3], Suvarna Chaure[4]

[1,2,3]Student, Dept. of Computer Engineering, SIESGST, Nerul, Maharashtra, India
[4]Professor, Dept. of Computer Engineering, SIESGST, Nerul, Maharashtra, India

---

Abstract - *Nowadays, the amount of people using social media is very huge and has been increasing day by day. However, due to increasing usage of social media platforms like Facebook, Twitter, Code mixing has also been commonly used in verbal communication. Code Mixed language like Hinglish, an informal language which is combination of Hindi and English is widely accepted in India as people feel more comfortable speaking in their own language. Thus, this growing practice of multi-lingual or code-mixed data is becoming a mainstream approach. So this paper provides an approach for the analysis and classification of code mixed language using Wordnet technique. We segregate our words in sentence to English and Not English words. After which we process the English words separately and calculate the polarity contributed by English words. We even process the Not English words separately and calculate its polarity. Then the overall polarity of that sentence is determined by combining the polarities of English and Not English words. Hence with experimental assessment using code-mixed data sets and classification using Polarity Detection, where Text can be classified positive polarity or negative polarity using polarity scores.*

***Key Words*: Hinglish, Not English, Wordnet, Code-mixed, Polarity, Devanagiri.**

1. INTRODUCTION

Today, the Rapid evolution of social media has created many new opportunities for information access and language technology, and also every person is digitally connected through it. But it also has created many new challenges, making it today's prime research areas. Although English is still by far the most popularly used language in Social Media context, its dominance is already being challenged due to the increasing popularity of regional dialects amongst native users. However, due to increasing informal usage of local languages in social media platforms, multi-lingual or code-mixed data is rapidly becoming a common occurrence. Mixed code is something when users use more than one language.

For e.g. Chalo jaldi karo, or we'll miss the beginning of the movie[1]. Here, the sentence is a mix of English with Hindi or Hinglish as it is popularly known. This sentence has meaning in Hindi language but it is written using Roman English instead of Hindi Devanagri script. It is very important to analyze this type of text to identify the sentiments, likes and dislikes, preferences, experiences of people as most of these comments or texts appear to be in this format nowadays.

Today's world is so socialized that each and every person we see mostly has accounts on all the social media. We require a language for conversation that is easy to type and read and also to understand. Hence people mostly prefer the local languages. Since typing text in English roman script is easier than typing texts in Hindi Devanagari or Marathi Devanagari script or other languages, people use combination of both the languages. Hence we require to analyse the sentiments so as to understand the approach of people.

1.1 NEED OF THIS SYSTEM

To understand the customer's sentiments often companies store huge amounts of customer feedbacks, reviews, market trends to understand the user's needs, experiences, preferences, overall opinions and also to capture the latest trends going on. Social media platforms like Facebook, twitter and other social networking sites often analyse such data through posts, comments, likes etc. In the Country like India, which is the home to many languages. They use phonetic typing or roman script and frequently insert English and various words or phrases through code-mixing or often mix multiple languages to convey their thoughts and opinions.

So, such data provides a significant challenge for organizations to understand and identify the sentiments. So hence it is important to analyse such texts and data. Also, a negligible amount of work has been done in this domain. Therefore, our aim to find the sentiments of user behind the code-mixed data.

1.2 OBJECTIVE

Analysis of code-mixed data is useful in many fields. People on Social media can use to find the sentiments related to a content maybe a post, story, comment etc. It helps others to understand what a person is thinking about it. It in a way even helps to report a post if it is inappropriate. It is also useful in Recommendation systems. YouTube, Amazon, Flipkart, bookmyshow etc are used for recommending the various suggestions based on their recent search history, past history and their views, ratings and comments.

2. PROPOSED SYSTEM

The current systems shows that for Hindi and Hinglish text, almost all the work is done using Dictionary. It is a process that takes maximum time and does not guarantee results. There are millions of words used and as adding each word with associated meaning and polarity makes it not only time consuming but also limits the words within the Dictionary. We need to add words with its different forms along with the meaning which becomes a tedious job. Also addition of new word would decrease the success rate of algorithm.

The dictionary that is used here is the one used for knowing the meaning of the words. Words are manually added to Hinglish dictionary which contains Hinglish and another Dictionary containing English words. All possible and most frequently used words are added in the dictionary. In the dictionary, Hinglish words are converted to English synonym and English words are kept as it is[1]. Then text pre-processing techniques are used to eliminate the words that do not contribute to any sentiments. Later on algorithms are applied to conclude whether the sentence is negative or positive.

The proposed system focuses on classifying the sentiment using wordnet instead of using the traditional approach to sentimental analysis[3].

3. SYSTEM DESIGN

In code-mixed sentences which is the combination of Hindi and English language, the first step is to identify the language of each word. If it is English then it is tagged as /E and if the word is Hindi then /H[2]. The next step is to calculate the polarity of English words with Wordnet. For Hindi words, they are first translated to Devanagari Hindi and then their polarity is calculated using HindiWordnet[2].

3.1 WORKING BREAKDOWN

The working of the system is broadly divided into following sections:

1. Data Collection Phase:

   There are not much Hinglish dataset available readily because analysis of Hinglish text is not that popular hence we have collected sentences related to movie domain from various blogs and social media sites like Facebook, Twitter, Youtube, bookmyshow, newspaper etc.

2. Text Pre-processing:

   It transforms text into a more digestible form so that machine learning algorithms can perform better.

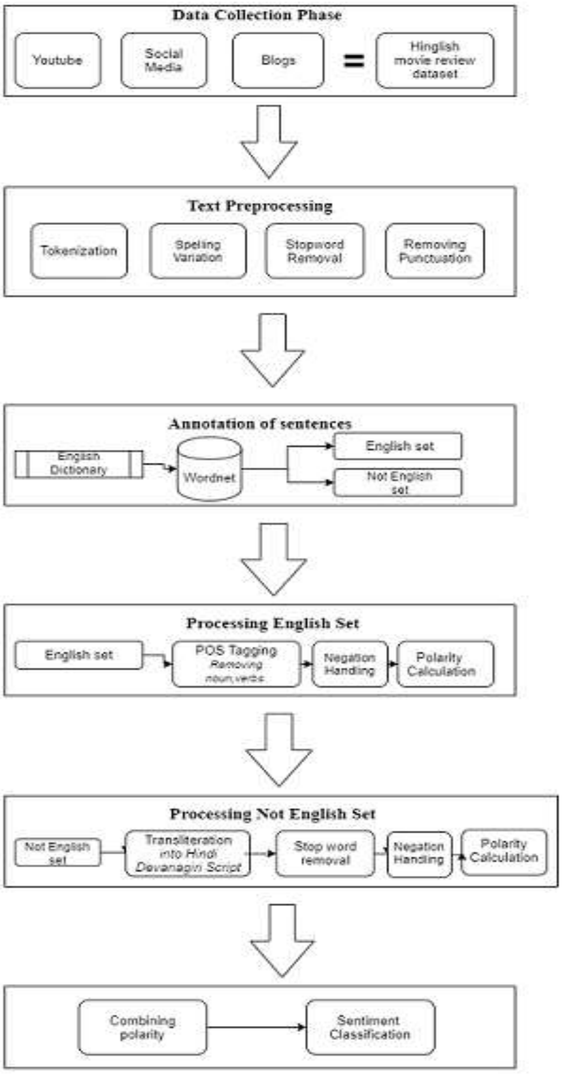

Fig -1: Proposed system for Hinglish text sentiment Classification

2.1. Tokenization:

It is the process of tokenizing or splitting a string, text into a list of tokens.

2.2. Spelling Variation:

It is very important to handle spelling variation because there might be some letters in the words which are wrongly repeated. Additional letters need to ripped off, otherwise they might add misleading information.
For eg: 'amazingggg' needs to be converted to 'amazing'. For this purpose spell correction algorithm is used which uses pattern.en library to handle spelling variations[4].

2.3. Stop word removal:

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) which do not contribute to any sentiment. These stop words are removed as it decreases accuracy using readily available English and Hinglish stop word list.

2.4. Removing Punctuation:

It is important to remove punctuations so that we don’t have different forms of the same word. If we don’t remove the punctuation, then been. and been, and been! Will be treated separately.

3. Annotation of sentences:

After text preprocessing, the remaining words are annotated as ‘English’ and ‘Not English’ by comparing it with English wordnet. If the word is present in the wordnet then that word is inserted in English set else it is inserted in Not English set.

The processing of English and Not English set is carried out separately.

4. Processing English set:

4.1. POS Tagging:

POS Tagging is done for the words present in the English list obtained from previous step. POS Tagging is the process in which each word is tagged to its respective part-of-speech and signifies whether the word is a noun, adjective, verb, and so on. Only the words tagged as adjective, adverb are kept for further polarity calculation and the rest are removed.

4.2. Negation Handling:

Negations are those words that affect the sentiment of other words in the sentence. For eg: In English negation words can be not, never, no etc. When these words occur in a sentence the polarity of that sentence obtained from adjectives, adverbs is reversed.

4.3. Polarity Calculation:

Polarity of each word obtained after POS Tagging is calculated using English synset. Synset in WordNet 3.0 is uniquely identified by 'POS & Synset#rank' pair which gives positive and negative scores for a word. Synsets are further divided into 'positive' and 'negative' classes depending upon the synset scores. A 'positive' label is assigned if the synset score is greater than 0 and a 'negative' label is assigned when the synset score is less than 0.

5. Processing Not English set:

5.1. Transliteration:

Words of Not English list are transliterated to its Devanagiri script so that the further processing can be done easily. Transliteration is done using indic trans. It helps to convert text from one indic script to another.

5.2. Stop Word Removal:

Once the Hinglish words are converted into its Devanagiri script, stop words can be easily removed by comparing with readily available Devanagiri hindi stop word list.

5.3. Negation Handling:

Similar to the negation words in English there can be negations in Hinglish also. These words can be nahi, agar, magar etc. Such words are handled by reversing the polarity of a sentence obtained using other words.

5.4. Polarity Calculation:

Polarity of the remaining words are calculated using Hindi sentiwordnet which has positive and negative scores for each of the word. If positive score>negative score then the word is labelled as ‘positive’ else the word is labelled as ‘negative’.

6. Combining Polarity:

This is the last step in which overall sentiment of a sentence is calculated by combining the polarity of all the words from English set and Not English set.

4. CONCLUSIONS

A very little work is done for Hinglish text analysis using Dictionary. There is no guarantee that it provides an accuracy of about 100 percent. Here in this system, we propose a method for analysis of Hinglish text using wordnet approach.

For this analysis, data set for Hinglish text is created manually from various reliable sources which contains sentences including both English and Hinglish words. Various text pre-processing techniques are applied using which only required words are kept for sentiment analysis. Then the words are annotated according to English or Not English and stored in two different sets and then the processing is done differently on both the sets.
Finally we combine the polarities from both the sets. Also negation is handled as it would reverse the polarity and meaning of the sentence.

Senti wordnet is the main tool for calculating the polarity. Also the spelling corrections are done in pre-processing stage. All unwanted words are also removed. Use of POS Taggers makes it easy to find the words contributing to sentiments. The main aim of this system is to efficiently analyse Hinglish text.

5. FUTURE WORK

Future work for this could be the use of smileys or emojis is also a form of expressing emotions and that should not be considered as noise. We can detect the emoji and compare with the emoji table having its meaning and sentiments[3].

Also punctuations marks can be considered and not pre-processed because that can show the sarcastic nature of the sentence. Rather than taking datasets online or manually creating them, we can take online reviews from users to enhance and improvise the efficiency and performance.

REFERENCES

[1] Harpreet Kaur, Nidhi and Dr. Veenu Managat, “Dictionary nased Sentiment Analysis of Hinglish text,” Volume 8, No.5, May-June 201, ISSN No. 0976-5697.

[2] Mohammed Arshad Ansari and Sharvari Govilkar, “Sentimental Analysis of codemixed data for transliterated Hindi and Marathi texts”, Volume 7, No.2, April 2018.

[3] Deepak Singh Tomar and Pankaj Sharma,”A text polarity analysis using senti wordnet”,Vol. 7(1), 2016, 190-193.

[4] Hitesh Parmar, Glory Shah and Sanjay Bhanderi, “Sentiment mining of Movie reviews using random forest with tuned hyperparameters”,