International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 12 | Dec 2021 www.irjet.net p-ISSN: 2395-0072

Sentiment Analysis of Code Mixed Text

Ritu Pawar[1], Swapna Salunkhe[2], Shilpa Dhongdi[3], Vaibhav Waghamare[4], Prof. Tushar A. Rane[5]

[1,2,3,4]Dept. Information Technology, SCTR'S Pune Institute of Computer Technology, Pune, Maharashtra, India,
[5]Assistant Professor, Dept. Information Technology, SCTR'S Pune Institute of Computer Technology, Pune, Maharashtra, India

---

Abstract - *In this covid-19 pandemic there is the huge inclination towards networking sites and social media asnow people cannot gather therefore the only mode of communication is via digital technology. Due to this it has become lot easier for government as well as variousindustries such as tele-marketing, commercial etc to analyze the opinion of people towards any social issues or reviews for any newly launched products, on which they can further take the required actions.However, to do so, the respective body needed to decrypt the content which is being posted by the people which isusually in form of mixed complicated combination of multiple language. Nowadays people post their opinions and reviews generally in the combination of Romanized English with their own mother tongue. In this paper we studied various Machine Learning approaches and proposed efficient techniques to identify the sentiment of the text after normalizing and predicted those sentiments in various categories.*

***Key Words*: Machine learning, Sentiments, Romanized English, Reviews, Code-mixed**

1.INTRODUCTION

With the progression of innovation and less expensive accessibility of electronic devices, pretty much every human is carefully associated through Social media. Also, as per the overview, the complete time spent viaweb-based media in the world across PC and cell phones expanded by 83% . Online Media has gone past basically friendly sharing to building notoriety and acquiring vocation openings. Associations are utilizing web-based media as a device to comprehend their clients, items and administrations and these audits are not just restricted in one language. These feelings are best owed in a combination of dialects.

This pattern of blending two dialects in talking and composing has prompted instituting of two terms: code-exchanging and code mixing where Code- Switching is "the juxtaposition inside the same discourse trade of sections of discourse having a place with two unique linguistic frameworks or sub- frameworks" (Gumperz 1982), and Code-Mixing alludes to the inserting of etymological units like expressions, words and morphemes of one language into an expression of another dialect (Myers-Scotton 1993, 2002). Hence, Code-Switching is normallybetween sentences while Code-Mixing (CM) is an intra- sentential peculiarity. Code exchanging is generally found during casual correspondence like websites,tweets and posts.

In Indian Online media, we found blending of Hindi and English language. Language variety triggers Indians to habitually change also, blend dialects. This peculiarity of code-blending has been seen by bilingual speakers from metropolitan urban areas where the individual talks in English at the work spot and Hindi or other Indian Languages at his home. Furthermore, the other explanation, why individuals blend two dialects whiletalking is that they are not really familiar with their native language dialects or they have a greately improved and less complex substitute of a word in the English language. Presently a-days, we additionally track down a high measure of code-blending between Hindi-English language in Bollywood film exchanges and melody verses.

Today, practically well known bollywood melodies have Hinglish verses. It makes the melodies appealing and extremely engaging and the crowds love them. This code-blend script is written in Roman letters in order. Here, the words from one language are phonetically written in other language utilizing Romancontent. These phonetically spelled words don't have astandard spelling; so for each word we find a few spellings. The text accessible on Social media has a great deal of deviations from the standard orthography, (for example, Word play, Creative spellings and so on) Clients utilize non-standard contractions and non–etymological sounds.

The attributes of the language viewed as on online media are portrayed by Danetand Herring, 2007. We additionally find a ton of utilization of "smileys", which pack an articulation into one or hardly any symbols. These realities represent a significant challenge in recognizing the language of the text, Information Extraction, Machine Translation and SentimentAnalysis. An illustration of an assertion in English would be: "Sibling, you are incredible!" A similar assertion in unadulterated Hindi would be: " , " Nonetheless, via web-based media, such a proclamation would be composed like this: "Bhai, tu extraordinary hai !" This is an example of Code-blend.The assertion is a combination of two dialects. Words Bhai, tu, hai – have a place with the Hindi language and the word incredible – has a place with English language.In this examination we have zeroed in on recognizing the language of code blend script (English and Hindi) which likewise included taking care of the different related qualities of code-blend.

2. LITERATURE SURVEY

Shashank et al. (Dec,2015) in a paper sentiment analysis code- mix script had done analysis of Hindi and English mixed language since Hindi is mostly combined with English language[8]. Author proposedapproach in which they divided system into 2 parts. First language identification with English dictionary and to identify Hindi words they used token from corpus of Hindi lyrics. In second part they done sentiment analysis of sentence using 3 sentiment resources - Opinion lexicon, AFINN (affective lexicon) and lastly WordNet (combination of dictionary and contains 155287 words). Since WordNet was giving better results compared to others so they used WordNet API. In that approach they obtained 80% accuracy.However, in this approach they failed to classify someof sentence. For e.g. "Movie itne achche nai lagi, bhai !! fine hai bus."Is negative sentence but the words "fine" "achche" is positive so the result given by system is positive.

Sreelakshmi k, Premjith B, Soman K.P(April,2020), conducted research to detect hate speech text in Hindi English code-mixed data using various machine learning models.The proposed methodology makes use of Facebook's word embedding pre-trained library, fastText to represent across 10000 samples of data collected from various sources as non-hate and hate. The study provides an insight to researchers working in this fieldof code-mixed data that the best result is being provided at the character level.The performance of the proposed system is compared with doc2vec and word2vec and it is being observed that fastText features gives better representation withSVM-RBF classifier. Thus the study confirms that proposed approach has performed well compared to all existing work done in this area but in future it could be extended to finer classification according to the percent of hate speech detected and this approach could be used in classification of sentiments depends on the hate speech, if it contains more offensive word that could be classified as negative else it could be positive sentiment.

Kaur, Harpreet, Veenu Mangat, and Nidhi Krail (May,2017) conducted a on Sentiment analysis of the "Hinglish" text using the dictionary based technique which was proposed in [3]. Movie review dataset in "Hinglish" was collected for analyzing sentiment. The authors here prepared two dictionaries: one is for English data, and another is for Hindi data which has the ability of handling word variations and also are case-insensitive. Tf-idf technique was used along with unigram, bigram and trigram for feature extraction, For sentiment classification, Naïve Bayes, SVM, Logistic regression and Neural Network algorithms were used for identification of bests feature set and also the classifier for "Hinglish" text.

Geetika Gautam et al. (sept,2014) In the research work of a set of techniques with semantic analysis of machine learning to classify the sentence and reviews of product that was based on twitter data[4]. The mainaim of the research work was to analyze a big amountof reviews (Twitters) by using twitter dataset that were already labelled. Further the accuracy was calculated and there was improvement of 1.7% in thatwhen the semantic analysis WorldNet was followed up and taking it to 89.9% from 88.2%.

Abinash Tripathy et al. (May,2015) proposed that sentiment analysis is the most well-known branch of natural language processing. To identify the intention of the author it deals with the text classification of text.The aim is of appreciation (positive) or criticism (Negative) type. The proposed work also shown aevolution analysis of results gaining through classification methods Naive Bayes and Support Vector Machine. These classification methods were used for classification intension in a sentimental review havingeither a appreciation (positive) or criticism (Negative)review.

3. TECHNIQUES

1] MACHINE LEARNING APPROACH

Machine learning approach uses supervised, semi-supervised or unsupervised learning to construct a model from a large training corpus. As for this approaches, supervised learning techniques have been the most widely used in many sentiment analysis tasks. It makes use of a training corpus to learn a certain classifier function. Using supervised algorithms the efficiency of sentiment analysis systems depends on the combination of appropriate algorithms together with a set of appropriate features.

Machine Learning techniques uses a labelled training dataset to train a classifier. Most machine learning methods treat sentiment analysis as a supervised learning problem, though recent methods have explored semi supervised approaches as well.Unsupervised approaches are difficult to implement, because they require a huge amount of training data to be for accuracy. Furthermore, unsupervised methods may not always match up with human conclusions regarding the given text. Considering sentiment analysis as a classification problem, supervised techniques are more suitable. But availability of plenty of unlabeled data makes it worthwhile to peruse unsupervised methods as well.

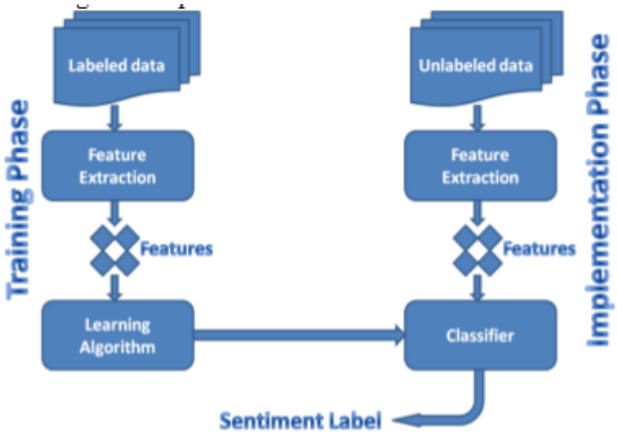

Fig -1: Architecture diagram

Among the many machine learning algorithms K- Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Maximum Entropy (ME) and Naive Byes (NB) classification techniques have been applied to classify the sentiment and evaluate the performance of the classifiers.

2] MAPPING BASED MODEL

In this approach, first independently learn monolingual word representations, L1 and L2, from large monolingual corpus and then learn a transformation matrix W to map representation from one language to the representation of the other language. Vector space of embedding of one language, L1 to that of the other language, L2 can be transformed using projection matrix. In both supervised and unsupervised settings,the projection matrix can be learnt. In an unsupervised setting the mapping learnt can perform better to those learnt in supervised settings. MUSE can be used to obtain this mapping. This tool returns a word aligned embedding space for each language such that similar words in either language have similar representation. By taking an average of word vectors of common words and simply appending the others two differentembedding spaces are merged. A single embedding space, L, and vocabulary for both the languages is obtained as a result of this.

MUSE provides both supervised and unsupervised variant.

1) MUSE-Unsupervised

To learn the transformation matrix followed by refinement using Procrustes algorithm it uses adversarial learning. Therefore, this approach is independent of any kind supervised signals between two languages.

2) MUSE-Supervised

It makes use of bilingual dictionary to learn a mappingfrom the source to the target space using Procrustes alignment. It is supervised approach as it requires bilingual dictionary.

3] PSEUDO-MULTI-LINGUAL

In this approach, we obtain word representation by training word embedding model, like fastText or Glove, on single corpus containing text from both the languages usually formed by concatenation of monolingual corpus. With this formation of a shared vocabulary and all words vectors are in a common vector space is done. Words from the two languages may or maybe not be aligned. Here subword based word embedding models are chosen. To guess the meaning of out-of-vocabulary words, which helps us tohandle the large vocabulary and large amount unseenconstructions caused by combining lexicon and syntax of two languages subwords are used. There is a huge variations of spelling and misspelling because most of the code-mixed text belongs to the category of texting language. Proper word vectors are obtained for even misspelled words using subwords.

FastText is popular method for learning word embeddings. It can create word vector for words absent from the trained embedding space, by summing up the character n-grams vectors. Therefore, it offer advantages similar to that of subwords.

4] RULE OR LEXICON BASED APPROACH MODEL

In code mix text (English + Hinglish), the very first step is to figure out language of the word. If its English then it is being tagged as /E and if it is Hindi then as /H. There after sentiment polarity of English letters is calculated with the help of Wordnet. For Hindi words,they are first converted to Devanagari Hindi and then sentiment polarity of those words is determined using Hindi Wordnet. It counts the total number of positive and negative letters in the given text. If the number of positive words is more than that of the negatives, it will return an answers as positive sentiment. If both are equal, it will then return a neutral sentiment.

The drawback of this method is that it does not take into consideration how the words are blended in a sentence, it only looks at incidents. It is fast to implement but the model requires a long-term cost outlay as it usually requires regular maintenance so that you get steady and enhanced results.

4. PROPOSED METHODOLOGY

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

1. Data Collection Phase:

There is not much Hinglish dataset available readily since analysis of Hinglish text isn't that popular thus we have collected most of sentences related to movie domain from numerous of blogs and social media siteslike Twitter, Facebook, Youtube, bookmyshow,newspaper etc.

2. Text Pre-processing:

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

2.1 Tokenization:

It is the procedure of separating a string, textinto a list of tokens.

2.2 Spelling Variant:

It is very essential to handle spelling alternatives because there might be some characters in the words which are erroneously repeated. Supplementary lettersneed to run off, otherwise they may add false info. Foreg: 'goooooooooood' needs to be converted to 'good'. For this purpose spell correction algorithm is used which uses pattern.en library to handle spelling variations.

2.3 Removing Punctuation:

It is essential to get rid of punctuations so we don't have distinct forms of the same word. If we don't cut off the punctuation, been and then been, and been! Will be considered safely.

2.4 Handling Phonetic Typing:

A rendered Hindi word can have numerous spellings and these spellings fluctuate due to the mother tongue effect. For example, श (sha) is marked as स (sa) by an Odiya person. And the phrase भीतर could be written as bhetar, betar ,bitar, vitar, bhitar. To handle these variants in the spelling, we have plotted a few uniformities to the same words.

2.5 Wrong Spelling Removal:

If we must classify the word 'reccomend', which has been mistakenly spelt, and the genuine spellings are 'recommend'. So we first of all be concerned about thisword and calculate it's Levenshtein distance.

2.6 Removing Repetitions of Word

It is necessary to eliminate described words as user usethose words constantly to emphasize on it but while managing single word will convey the adequate knowledge to the user and will be simple to process.

3. Translation:

After text preprocessing, the whole sentence is converted into English via google api translator.

4. Sentiment classification using Deep learning algorithm

Deep learning is a study within machine learning that uses "artificial neural networks" to process information much like the human brain does.

Deep learning is hierarchical machine learning that uses multiple algorithms in a progressive chain of events to solve complex problems and allows you to tackle massive amounts of data, accurately and with very little human interaction.

Recently, deep learning algorithm delivered spectacular performance in information science applications encompassing SA across various datasets. Such models wouldn't like any pre-defined options, however they may learn sophisticated options as of the dataset by themselves. though each single unit in theseNeural Networks (NN) is easy, by suggests that of stacking layers of NL units at the back of 1 another, those models area unit competent to learn extremely refined call boundaries. Words are signified in an exceedingly high-dimension vector area, and therefore the feature extortion is left to the NN. Architectures like RNNs also are competent to effectively comprehend the sentences structure. These build deep models the simplest work for tasks like sentiment analysis.

5. SYSTEM ARCHITECTURE/WORKFLOW

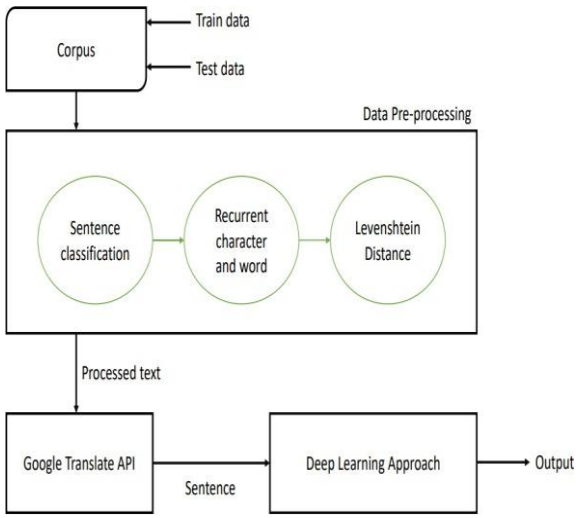

Fig -2: Workflow of system

6. UML DIAGRAMS

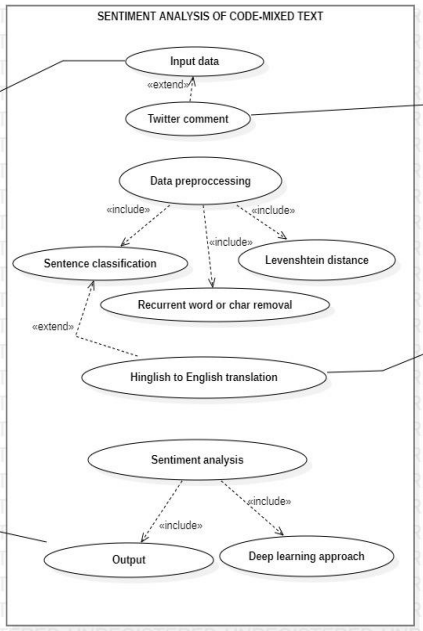

Fig -3: Use case diagram

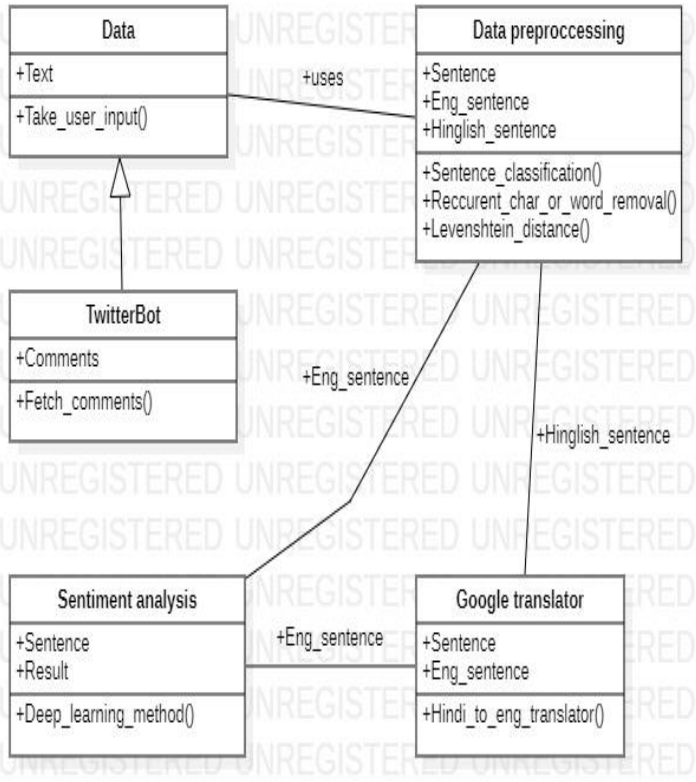

Fig -4: Class diagram

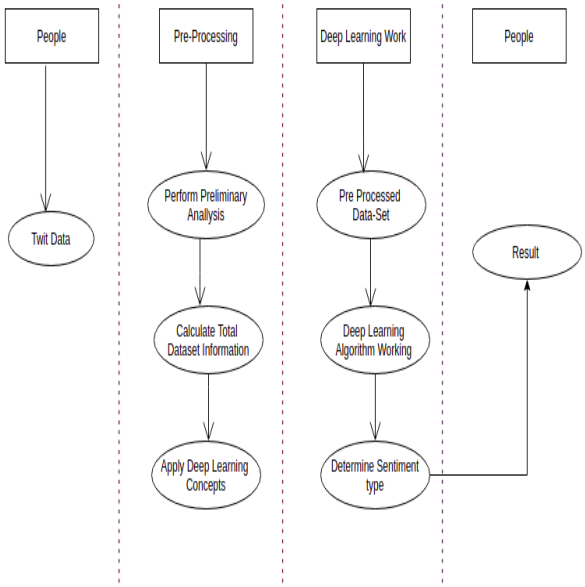

Fig -5: Activity Diagram

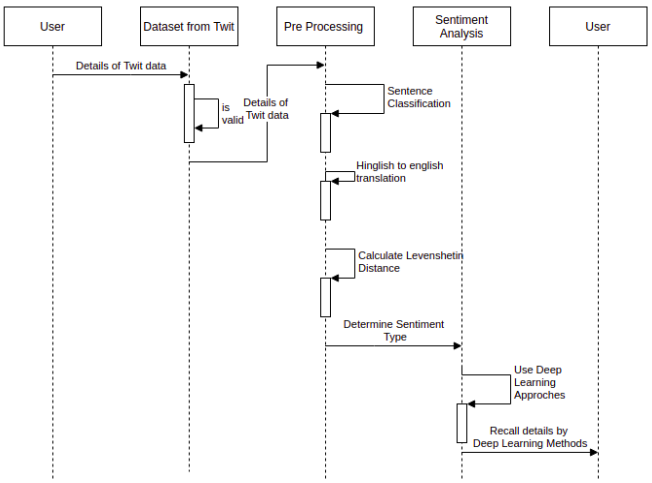

Fig -6: Sequence Diagram

7. CONCLUSION

In our work, we had the option to arrange a code blend explanation into positive or negative. This undertaking included recognizing the dialects (Hindi or English) in a code blend articulation, transcribing the Hindi Romanized script into Devanagari content, and afterward passing judgment on the feeling of the explanation. We expect this work will empower a superior comprehension and investigation of Social Media opinion in client discussions across the world and particularly in Social Organizations in India. We are intending to stretch out our work to find and distinguish feelings in discussions on friendly organizing locales and applications.

8. FUTURE SCOPE

Future scopes that need to be solved like ambiguity problems and aspect-based sentiment analysis in text.Detect Fake and spam review further research is expected to improve accuracy in sentiment analysis ofcode-mixed language.

Also, we would like to extend our work to several other language pairs of code-mixed data. It would be interesting to utilize the rich features of individual languages to help identifying sentiments in their code-mixed version.

REFERENCES

[1] Kaur, Harpreet, Veenu Mangat, and Nidhi Krail. "Dictionary based Sentiment Analysis of Hinglish text." International Journal of Advanced Research in Computer Science 8, no. 5 (2017).

[2] Detection of hate speech text in hindi English code mixed data ,third international conference on computing and network communication published by Elsevier B.V

[3] Harpreet kaur ,Dictionary based Sentiment Analysis of Hinglish text International Journal of Advanced Research in Computer Science

[4] Geetika Gautam Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis ,7th International Conference on Contemporary Computing, At: Noida(India) Volume: IEEE Xplore

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

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

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

[8] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques" Procedia Computer Science

[9] Sreelakshmi k, premjith B, soman K ,"Detection of Hate speech text in hindi-english code mixed data". Conference on Computing and Network Communications (CoCoNet'15).

[10] Shashank Sharma, Sentiment Analysis of Code – Mix Script 2015 Intl. Conference on Computing and Network Communications (CoCoNet'15), Dec. 16-19, 2015, Trivandrum, India