Sentiment Analysis of Hinglish Text

**Abstract:-**

In order to do sentiment analysis of Hinglish movie reviews commonly used in India, we tried different techniques to pre-process the reviews and a number of classifiers to

**Introduction:**

Sentiment analysis is a type of data mining that measures the inclination of people’s opinions through natural language processing (NLP), computational linguistics and text analysis, which are used to extract and analyze subjective information from the Web - mostly social media and similar sources. The analyzed data quantifies the general public's sentiments or reactions toward certain products, people or ideas and reveal the contextual polarity of the information.

Sentiment analysis helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences.

As per comprehension review of Ravi and Ravi [1], most of the work in this field has been done on text written in pure English language and a few work is done in other languages like Arabic, Spanish, Chinese etc. The amount work done on Indian languages further decreases

In this paper we did the sentiment analysis of one of the most used languages in Indian context i.e. Hinglish, which is written entirely in Latin script but containing informal words, phrases or even slang in a piece of writing, written both in Hindi and English.

For instance, consider a movie review written in latin text but Hindi language:

“gaane philm mein thoonse gae hain”

To accompany sentiment classification (whether positive or negative) with such text, we applied bag of words approach. For pure English text, work has already been done but for hinglish everything is needed to be done from scratch. Like while creating sparse matrix, the main problem is to match same Hindi words written with different spellings, e.g., “paisa” and “pesa” both have the same meaning “money” in English, but can be written in different ways by different users. To counter this problem we use technique of Lavenstein method to find most similar words and finally used Soundex function to give the words a score a score based on their pronunciation. If the difference between the new word and to be matched word is lesser than certain value then both words are supposed to be same else different column is made in sparse matrix. A dictionary of hinglish words is also used in the model to improve the accuracy of the approach.

For classification purpose, we used various classification techniques like SVM, logistic regression, Random Forest, Naive Bayes, Decision Tree, etc.

The contribution of this survey is significant for many reasons. This survey can be useful for e-commerce companies, movie productions and social media content analysis as it covers analyzing precious customer’s feedback written in different language which cannot be analyzed using traditional methods.

Literature Survey:

Since our work completely depends on machine learning approach so we did our literature survey in sentiment analysis using machine learning algorithms.

Richa et al [5]

[ Sharma, R., Nigam, S., and Jain, R. Polarity Detection of Movie Reviews in Hindi Language. International Journal on Computational Science & Applications 4, 4 (2014), 49-57 ]

applied dictionary-based approach to analyze the polarity of review. In this work, result is the collection of reviews among positive, negative and neutral reviews of the sentiment of the sentence.

Kumar et al. [6]

[ Ravi, K. and Ravi, V. Sentiment classification of Hinglish text. 3rd IEEE International Conference on Recent Advances in Information Technology (RAIT) (2016) ]

experimented with different combinations of feature selection methods and a host of classifiers using term frequency-inverse document frequency feature representation**.** They carried out in total 840 experiments in order to determine the best classifiers for sentiment expressed in the news and Facebook comments written in Hinglish. The experiment concluded that a triumvirate of term frequency-inverse document frequency-based feature representation, gain ratio based feature selection, and Radial Basis Function Neural Network as the best combination to classify sentiment expressed in the Hinglish text.

Pooja et al [8]

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

applied Hindi SentiWordnet (HSWN) to find the polarity of reviews made on Hindi movie. They performed document-level sentiment analysis. The polarity of words in the reviews is fetched from HSWN and then aggregated to find the overall polarity of the review . This system has two objectives: 1. Improving existing HSWN Wordnet. 2. Sentiment extraction. For the first objective, English SentiWordnet is used. Words which are not in HSWN are first translated into English and then their polarity scores are found. In the second objective, the sentiment is extracted by finding the overall polarity of the document that can be positive or negative or neutral.

Deepali et all

[ Mishra, D., Venugopalan, M. and Gupta, D. Context Specific Lexicon for Hindi Reviews. Procedia Computer Science 93, (2016), 554-563. ]

develop an improvised polarity lexicon to overcome the limitation of existing HindiWordnet which is very generic in nature. They work on datasets of hotel and movie domain. The built lexicon reflects context sensitivity and shows an improvement in the accuracy. Their system has two objectives: To build an improvised context sensitive polarity lexicon for a specific domain and to improve the lexicon coverage by using the approach of HindiWordnet. Preprocessing involves tokenization, POS tagging, and lemmatization i.e. reduction to root word (For example ‘lovely’ or ‘loved’ can be reduced to ‘love’). Then tf-idf score of each opinion word is calculated and words whose tf/idf scores are above some predefined threshold are kept and rest are discarded. Next step is to calculate final polarity score. This is termed as Context specific polarity lexicon (CSPL). Then adverbs and adjectives are extracted from this lexicon. Their synonyms are extracted from HindiWordnet. If a synonym is present in CSPL then there is no change in polarity score, otherwise extracted synonym is added to CSPL and same polarity score is assigned to it.

Shanshank et al [9]

[ Sharma, S., P. Y. K. L., S. and Rakesh Chandra, B. Sentiment analysis of code-mix script. IEEE International Conference on Computing and Network Communications (2015). ]

perform sentiment analysis of code mix statements i.e. statements having both English and Hindi sentences. They use a statistical method to find the sentiment. In this, if a number of positive words in the statement is more, then the statement is termed as positive and vice versa. The methodology of this system involves language identification which involves tagging the words to corresponding languages as tag /E for English and /H for Hindi. After this spelling correction is performed e.g. happppyyy becomes happy. Then some words which are present in both languages (Hindi and English) like ‘so’, ‘do’, ‘teen’ etc. are handled in ambiguous word handling phase. Sounds like ‘awww’, ‘boo’, ‘opps’ are also taken into consideration. Phonetic variations are also handled. Then, Roman Hindi is transliterated to Devanagari Hindi. In the end, Wordnet is used for finding the polarity of English and HindiWordnet is used for finding the polarity of Hindi words.

Prashasti et al.[10]

[ Kanikar, P., Koppisetty, R., Govindan, S., Bhat, S. and Virani, M. Semantic Analysis on Twitter Data Generated by Indian Users. International journal of scientific and engineering research 7, 9, (2016)]

perform semantic analysis of Hinglish tweets. They use a dictionary based approach. For determining the polarity of tweets, a Scoring algorithm is used. Naïve Bayes classifier is used for classification purpose. They built two dictionaries one for positive words and other for negative words. The dictionaries contain all the words from tweets and also polarity scores associated with them. Words in tweets are matched with dictionary words.

Rupal et al [13]

[ Bhargava, R., Sharma, Y. and Sharma, S. Sentiment analysis for mixed script Indic sentences. IEEE International Conference on Advances in Computing, Communications, and Informatics]

worked on multilingual sentiment analysis. The code mix consists of English and four other Indian languages (Tamil, Telugu, Hindi and Bengali). Multi-class SVM is used for classification. They use dictionary-based approach for training data for all the languages. They have taken FIRE 2015 dataset. Training dataset was built by scrapping different websites like bengalilyrics.com, Tamil lyrics, and other Indian language websites. The first task is to identify the language and tag it. Indian languages are transliterated to their original form. N-gram features are extracted and classification is performed using multiclass SVM.

**Proposed Approach:**