International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 2, ISSN No. 2455-2143, Pages 245-250
Published Online June 2021 in IJEAST (http://www.ijeast.com)

SENTIMIX (HINDI- ENGLISH)

Aastha Awasthi
Department of IT
IMS Engineering College, Ghaziabad, UP, India

Akanksha Singh
Department of IT
IMS Engineering College, Ghaziabad, UP, India

Istuti Agarwal
Department of IT
IMS Engineering College, Ghaziabad, UP, India

Arushi Sanjay
Department of IT
IMS Engineering College, Ghaziabad, UP, India

Abstract— **The study of emotions is an important part of research in the excavation of literature. It involves removing ideas from articles such as reviews, news, blog posts, etc. and then classifying them as positive, negative or negative. The senses of English were studied but not much work was done for the Indian language. Studies were conducted in Hindi, Bengali, Marathi and Punjabi. Today, most communications are made on social media using the Hinglish language which is a combination of both Hindi and English. Hinglish is a native language that is very popular in India because people are more comfortable speaking their own language. This paper provides a new method for expressing the emotions of Hinglish (Hindi + English). Emotional analysis (SA) uses mixed data from the social media with many programs. Feedback from consumer satisfaction in evaluating social activity in multiple languages society. Progress in this area is hampered by a lack of relevant sequential data. We Introduces a Hi-En code-mixing set for sensitive information and satisfying performance. A comparative study of the feasibility and implementation of SA methods in social media. We also derive and describe Hinglish language. We examine this problem by using sets of lexicon, emotion, and form metadata to construct a classification that can vary between "positive", "negative" and "neutral" feelings.**

Keywords—SVM, Random Forest, Django Framework, Codemixed, Polarity Detection, opinion mining.

I. INTRODUCTION

The word "sentiment analysis" and "opinion mining" is used interchangeably in this paper. The popularity and opportunity-rich content includes movie reviews, product reviews, blogs and posts. Emotions can be expressed in three levels: Data, data level and area / critical area. Polarity is calculated for the number of papers in the Document level Sentiment Analysis. The polarity is selected for the aspect / shape of the paper in the Aspect level Sentiment Analysis. The intention is to motivate the writer's mind to seek ideas related to the idea to a subject in the paper. Thus, a combination of expert and technical expertise for analytical writing and categorizing user emotions in positive, negative and negative classes. With the rapid evolvement of a medium of communication, developing countries like India, China, etc. are switching towards online media swiftly. Among various languages spoken in India, Hindi and english is one of the spoken languages in the world.

In addition to the problem definition, any man-made phrase has multiple meanings. People express their ideas in different ways; Rhetorical devices such as sarcasm, insults and gestures may distort the sense of information. The only way to truly understand these mechanisms is the approach: knowing how starting a paragraph can affect the sentiment of other internal sentences. To address the issue, much of the research surrounding theories has focused on the use of feature engineering. Constructing accuracy in a model that takes into account the background, voice, and past of emotions can help increase accuracy and lead to a more general understanding of what the author is trying to convey.

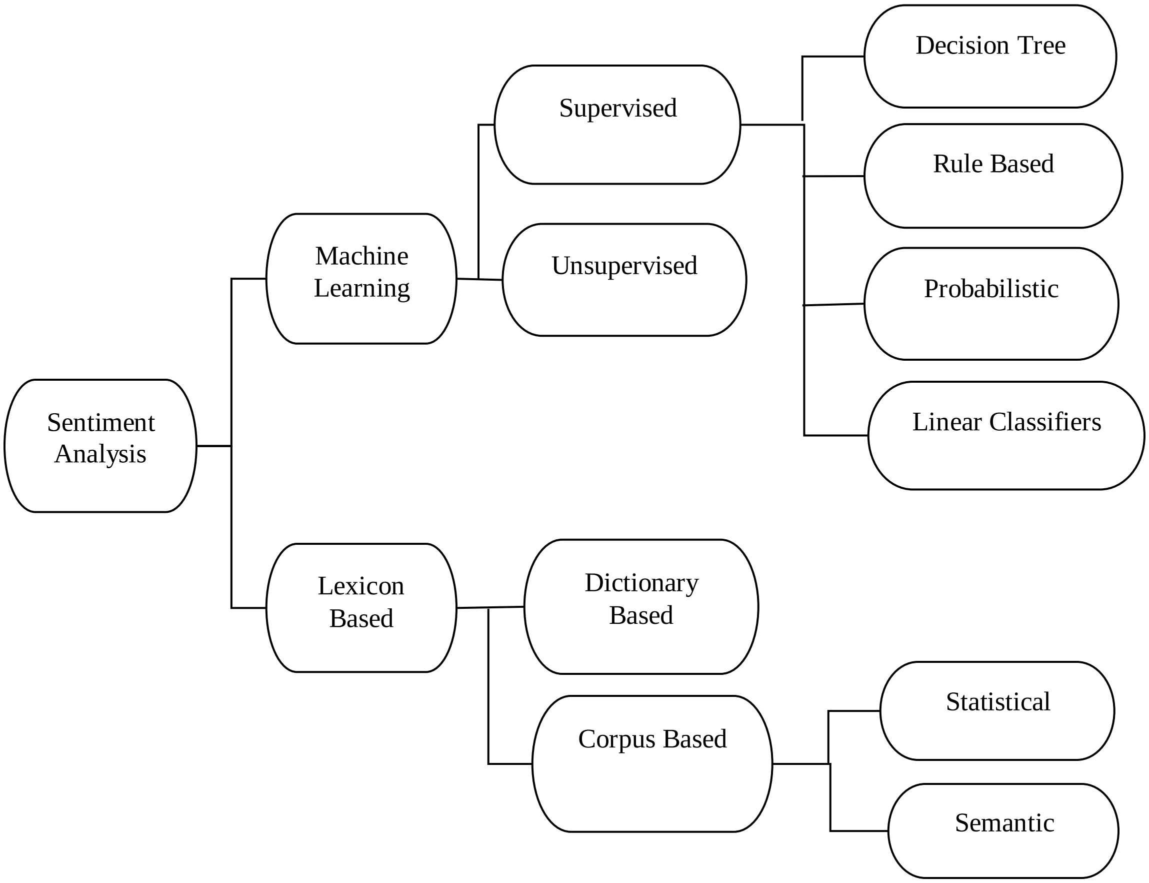
Sentiment Analysis, aka Opinion Mining, enough popular in and around education and industry for the latter one half a decade. Sensory support is a computational study feelings, emotions and attitudes towards a company In solid text or text communication, to express emotions. For details, we need to perform various functions such as subjectivity Classification, Emotional Classification, Aspect Removal, Spam Detection Ideas, Review Measurement Help, and more. Less research is also available In other languages such as Chinese, Turkish, Arabic, in Spanish and other texts. If we think about the signs Language -based studies for the study of emotions, however, are left behind compare with other languages. To fill this space, We planned to classify the emotions expressed in non –English Basic written language i.e Hinglish. Hinglish is a common language. Language, it is usually spoken by Hindi people who speak Hindi. Hindi words in Hinglish language as well as English in the same verse. There are two approaches for sentiment extraction: Machine learning and Lexicon-Based approach. Machine learning algorithms are mostly used for extracting sentiments at sentence and doc. levels. Naïve Bayes, SVM, and Maximum entropy are best supervised machine learning algorithms. Machine learning algorithms are used widely. Lexicon-based approaches to use a dictionary for sentiment analysis.

Let's think a Samples from our set,

"Ho Amir ho koi bhi ho law sabke liye ek jaisa hai …..so don't support Ram"(whether Amir or anyone else, the law should be the same for everyone ...so don't support Ram.)

The Hindi words of the sentence are Ho, Koi, Bhi, sabke , liye, ke, ek , jaisa and hai which are Mixed with English words to complete a sentence.

To the best of our knowledge, we didn't find any study in sentiment analysis, which is performed on Hinglish corpora. Therefore, we performed sentiment classification of Hinglish text written in Roman script. In order to determine the sentiment polarity of a text, we employed a task of sentiment analysis, called sentiment classification (SC). In order to classify sentiment expressed in Hinglish, we employed random forest, SVM classifer with linear kernel on the text obtained on the pre processing.



II. LITERATURE SURVEY

The current study is based on machine learning based approach; therefore we carried out literature survey on machine learning and Hybrid based approaches. We also surveyed some sentiment analysis studies carried out on code-mixed languages.

In the paper by (Alexander Pak, Patrick Paroubek) [1]includes n-gram as a binary feature, while for general information retrieval purposes, Tokenization

– they segment text by splitting it by spaces and punctuation marks, and form a bag of words and Constructing n-grams – they make a set of n-grams out of consecutive words. Conditional Random Field (CRF). Naive Bayes Classifier, SVM with four Kernels are used for classification to determine positive, negative and neutral sentiments of documents that uses N-gram and POS-tags as features.

[2]Harpreetkaur , Nidhi , Dr. Veenu Mangat proposed a dictionary based approach for sentiment analysis of Hinglish text using different ML techniques. To train the data, feature extraction
technique is used. For feature extraction TF-IDF is used along with unigram, analysis of Hinglish text using different ML techniques. To train the data, feature extraction technique is used. For feature extraction TF-IDF is used along with unigram, bigram, and trigram. Negation handling is also done.TF-IDF is the best feature extraction technique. Analyzing the accuracy with different classification techniques. The work which has been done for Hinglish text using translation is never 100% accurate and leads to decrease in accuracy.

Braja Gopal Patra, Dipankar Das, and AmitavaDas[3] used the Twitter API to collect both Bengali and Hindi code-mixed data from Twitter. It includes identification of Part-of-Speech (POS) tags in code-mixed Data. Different approaches which included features like Glove word embeddings and TF-IDF scores of word n-grams as well as character n-grams package is used to calculate the TF-IDF.LSTM is also used in Word2Vec.
Finally, two classifiers: ensemble voting which consists of three classifiers - linear SVM, logistic regression and random forests and linear SVM are used for classification. For calculation of positive sentiment, the predicted negative and neutral tags are converted to other to both gold and predicted output by doing binary classification. The team used word and character level n-grams as features and SVM for sentiment classification.

Pranay deep singh and Els Lefeve[4]r proposed the use of unsupervised cross-lingual embeddings for solving the problem of code-mixed social media text understanding. They specifically investigated the use of these embeddings for a sentiment analysis task for Hinglish Tweet. This paper investigates the use of unsupervised cross-lingual embeddings for solving the problem of code-mixed social media text understanding. They used the training data provided for the SemEval 2020 shared task on sentiment analysis in code-mixed social media text. As a result, the presented approach can be used for code-mixed text processing tasks in a variety of languages, and could be an important contribution to solve the data-acquisition bottleneck for NLP for codemixed data.

Gaurav Singh[5] , he proposed a method to predict the polarity of the social media text through machine learning. They use the model which consist of language identifying, conversion to English script and classificiation using SVM, logistic regression, etc. Dataset took from the twitter text, facebook comments other social media platforms. Here they find the polarity by classify the sentiment and keywords provided the tokens and find the language.
Logistic regression gave the best result followed SVM and random forest.

Varsha thakur, Roshni Sahu &Saumya Omer[6] aimed to classify the Hinglish Text using different Approaches i.e. Machine Learning and Lexicon Based. Made their own corpus (with help of 500 feedback from different social sites. In they collect hinglish data from two domain news and facebook comment). To train the data, feature extraction techniques are used which are Unigram, Bigram, n-gram Term frequency, Tf-Idf, POS tagging. Using SVM, Hybrid on Hinglish dataset (twitter domain) gets 96.8% accuracy. Ambiguity occurs due to not choosing right approach.

III. DATA

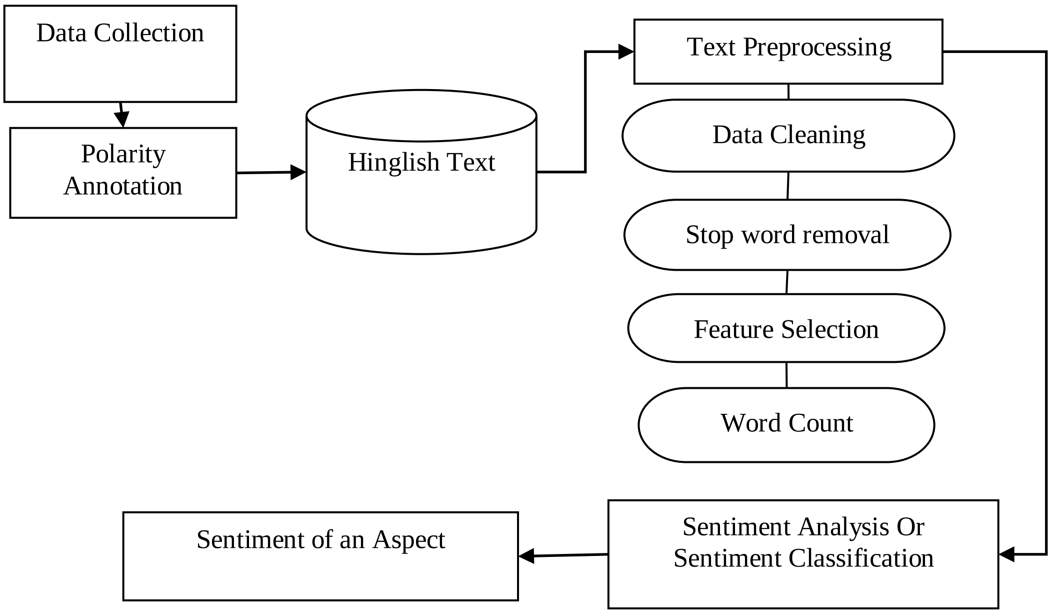
To train and evaluate our sentiment analysis system for Hinglish, we use the training data provided for the sentiment analysis in code-mixed social media or sentence given by the user. This dataset for Hinglish contains more than 6000 words, which have been labeled as positive, negative, or neutral.
For example:-

“Wow the weather is so amazing, chai ho jae ” which combines an English sentence with a Hindi sentence mid-way, Although the data set is tagged with a language label for every word, we did not use this information in our experiments as our aim was to build a common bilingual model that would be applicable for other code-mixed data sets as well.

IV. DATASET

The data was obtained from https://www.kaggle.com/mukulkirti/positive-and-negative-word-listrar and https://gist.github.com/mkulakowski2/4289441 and hindi words are added by ourselves. The data consisted of code mixed text containing hindi and english words written in english script. The text are classified among the Negative, Neutral or Positive sentiment polarity. The data is provided in .csv format.

V. MATERIALS AND METHODS



Sentiment Classification:

Just after the feature extraction, we get feature scores which are used to train the classifier. We have taken a number of classifiers such as SVM, Random Forest for experiments. Here, Our goal is to took out the best feature set and classifier for Hinglish text.

For example: To understand the procedure of Hinglish sentiment analysis, consider a review "yrr kya movie thi ..mazza aa gya…definitely enjoyed it ... everyone must watch it "

STEP1: Data Pre-Processing:

Data cleaning: Data cleaning is performed to improve the quality of data for modelling and then the results are taken. Custom methods for cleaning the data were made using regex and other predefined libraries in python. Data cleaning includes:

1. Removal of @ User, html tags were removed, html decoding was converted to symbols and special characters and also the removal of punctuation marks like Full Stop (.), Question Mark (?)Quotation Marks/Speech Marks (" "), Apostrophe ('), Comma (,), Hyphen (-) The dash (en dash (–) em dash (—)), Exclamation Mark (!),etc.
2. After the above process, removal of RT (Re-tweet keyword), https, nan (null character) keyword was done and then all the words were converted to lower case.
The given Hinglish Dataset of positive and negative words are going through the removal of special Characters and Conversion in lower case. In this Step all the special characters from the sentence are removed and conversion of all the upper case to lower case take place.

Creation of Stopword list: NLTK (Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages. Stopwords can be removed easily, by storing a list of words that can be considered to stop words. The list of

stopwords are:{‘ourselves’, ‘hers’, ‘between’, ‘yourself’, ‘but’, ‘again’, ‘there’, ‘once’, ‘about’, ‘during’, ‘very’, ‘out’ , ‘having’, ‘with’, ‘they’, ‘own’, ‘an’, ‘be’, ‘some’, ‘for’, ‘do’, ‘its’, ‘yours’, ‘such’, ‘into’,
‘of’, ‘most’, ‘itself’, ‘other’, ‘off’, ‘is’, ‘s’, ‘am’, ‘or’, ‘who’, ‘as’, ‘from’, etc.}

Stopword Removal: Stopword removal is an important preprocessing technique in text processing because the size of data can be reduced to 30-40%, without affecting its sentiments. NLTK allows removal of stopwords, and the list of stop words can be found in the corpus module. To remove stop words from a sentence, text should be divided into words and then remove the word if it exits in the list of stop words provided by NLTK.

Table 1: Examples Of Data After Pre-Processing.

SAMPLE TEXT WITH STOP WORDS	WITHOUT STOP WORDS
Can reading be exhausting ?	Reading , Exhausting
I like dancing , so i dance	Like , Dancing , Dance
I have no doubt that India will not win this time .	Doubt , India , win , time

STEP 2 : Word Count

In this stage the number of occurrences of each distinct word in the data set of both positive and negative reviews is calculated. We have created dataset of positive and negative words for both Hindi and English words. For positive words we have assigned 1 and for negative 0. As SVM classification or Random Forest classification uses integral values training dataset we have taken the words index as name and 0 or 1 for value. We train our model with these datasets. Then before prediction we extract words from sentence and take convert those words in integral form from our previous dataset. Now our classification model will predict the outcome.

DATASET TRAINING:

Support vector machine and random forest algorithm has been considered for classification in this experiment.

Support Vector Machines:
Support Vector Machines have been popular and effective models for multiclass classification problems. We used sklearn linear SVM library for implementing our SVM based models. We employed the one-versus-one strategy for the classification task. SVM is finding the best hyperplane that used to separates the data points of two different classes. For simplicity and because we are creating a sentiment tool, we call these classes ‘positive’ and ‘negative’, and plane represent a different class and this plane can thus be seen as a decision boundary for any new data point, because we can easily categorized this new point based on which side of plane it occurs. Therefore, after a decision boundary is obtained, a SVM is useful for making accurate predictions of new data points. The entire process of constructing a support vector machines can be divided into two parts. In the first part, we *train* a machine by providing it with a classified dataset. The training is complete by finding the best hyperplane that can separates the different classes. This hyperplane is denoted or represented by, where represents the dimensionality of the data. Training the machine thus implies finding the best values. In the second part, now we are able to use this trained machine to make predictions about the classification of any data point. For simplicity, we declare that positive points lies on the side of the plane for which it holds that and vice versa for negative points.

Random Forest:

Ensemble classification technique are learning algorithms which construct a bunch of classifiers as a substitute of one classifier and then classify new data points by taking a vote of their predictions. Random forest is a type of supervised machine learning algorithm which based on ensemble learning. Ensemble learning is that type of learning where you merge different types of algorithms or the same algorithm multiple times to form a more useful prediction model. The random forest algorithm combines or merges multiple algorithms of the same type that is multiple decision trees resulting in a forest of trees, hence the name "Random Forest"(RF). This algorithm can be used for both task regression and classification. RF classifier can be explained as the collection of tree-structured classifiers. Instead of splitting each node it uses the best split node among all variables, RF splits each node using the best among subset of predictors which randomly chosen at that node. A new training dataset is created from the original data set with substitution or replacement. After this a tree is grown using random feature selection.

WORD PREDICTION:

In this experiment we are using Random Forest and SVM at the same time for classification of text. Firstly, random forest is an ensemble learning method that construct multiple decision trees at randomly selected features and then predict the class of a test case by voting of the individual trees. Support Vector Machine turn around the concept of a margin- either side of a hyperplane that separates two classes. By maximizing themarginwe can create the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error. Random Forest is not sensitive to input parameters, thus, we just used the default parameters for each classifier. The trained classifiers returning scores between 0 and 1, these scores are then transformed to a binary state indicating negative or positive.

VI. RESULTS

The training dataset we used for this work were consisting of hinglish (Hn-En) words. The paper considered the supervised classification algorithms to analyze text. Support vector machine has higher performance with 90% accuracy than random forest with 80% accuracy. The best accuracy was given by the SVM. of Random Forest and SVM at the same time. Firstly, random forest is an ensemble learning method that construct a number of decision trees at randomly selected features and predict the class of a test instance by voting of the individual trees. Support Vector Machine revolves around the notion of a margin —either side of a hyperplane that separates two classes. Maximizing the margin and with this way creating the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error. RF was not sensitive to input parameters, thus, we just used the default parameters for each classifier. The trained classifiers return scores between 0 and 1, these scores are then transformed to a binary state indicating 'negative' or 'posi

VII. CONCLUSION

We reviewed three data sets to measure our support Vector Machine and Random forest problems. The first method we used for our problem was Random Forest. It is very easy and quick to train. In this system each character in each class thinks differently. The test is straightforward, sorting conditional probability from information available. One of the important tasks is to find the emotional posts that are so important in this process to get what is needed. In this inefficient way, we simply assumed that the terms were available in our state and set their specific terms. We have obtained some successful results after using this method on our problem. Secondly, we implemented for our problems supports core component analysis as well as vector machines. The main reason for using the main detail element is its reduced dimensionality. It is greatly reduced in size to small without any loss of information. The results of this approach are thus achievable. Work has been done for Hinglish texts using translations that are not 100% accurate, which lead to reduced accuracy. Here, we propose a way to implement the concepts of Hinglish writing using a methodological approach. In this we create two dataset: one for English records and one for Hindi records that can handle word variations. Hinglish stop word lists were also provided. We hope to find the best extracting and classification methods for Hinglish emotional writing However, we were able to predict emotions using different methods Support vector Machine random forest in machine learning. The best accuracy was given by the SVM.

VIII. REFERENCES

- Pruthwik Mishra, Prathyusha Danda, Pranav Dhakras: Code-Mixed Sentiment Analysis Using Machine Learning and Neural Network Approaches (2018).
- K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowledge-Based Systems 89 (2015).
- Harpreet kaur and Dr. Veenu Mangat : Dictionary based sentiment analysis of hinglish text(2017).
- Varsha Thakur, Roshani Sahu, Somya Omer: Current state of hinglish text of sentiment analysis.
- Gaurav Singh:- sentiment analysis code-mix of social media text.
- Brajra gopal, Dipankar Das, Amitva das: Sentiment Analysis of Code-Mixed Indian
- Language: An Overview of SAIL Code-Mixed Shared Task @ICON-2017
- Ameya Prabhu, Aditya Joshi ,Manish Srivastava : Towards Sub-Word Level
- Compositions for Sentiment Analysis of Hindi-English Code Mixed Text.
- Pranaydeep Singh and Els lefer[4]: Sentiment Analysis for Hinglish Code-mixed Tweets by means of Cross-lingual Word Embeddings.
- Alexander Pak, Patrick Paroubek :Twitter as a Corpus for Sentiment Analysis and Opinion Mining.
- P.V. Veena, M. Anand Kumar, Soman Kp: Character Embedding for Language Identification in Hindi-English Code-mixed Social Media Text.