IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 1, January 2023

DOI: 10.17148/IARJSET.2023.10130

HINGLISH SENTIMENT ANALYSIS

Rohan Singhal[1], Shrey Bishnoi[2]

CSE Department, Maharaja Agrasen Institute of Technology, India[1,2]

Abstract: One of the crucial fields in the contemporary technology world is sentiment analysis. Sentiment analysis, a branch of text mining, is the study of extracting feelings and emotions from actual data. Textual sentiment analysis can be helpful for many different decision-making procedures. Social media comments, on the other hand, frequently use scripts that are not their own and do not strictly adhere to grammar requirements. Hinglish, a language that combines Hindi and English, is widely used as aninformal writing language in India. based on aspects In the realm of artificial intelligence,sentiment analysis from Hinglish text is a difficult challenge to solve.

The most current advances in sentiment analysis from English text, code mixed text, andother issues associated with the same have been described in this work. The categorizationof aspect-based emotion is the primary difficulty faced by the Hinglish text model. Therefore, it is crucial to select the right categorization model. This study reviews prior work and describes many sorts of aspects of emotion text data and the extraction methodsrelated to them. Analyses of the reliability of several classification methods were also conducted. Additionally, language and textual characteristics of several ML approaches were analysed for actual word analysis.

I. INTRODUCTION

Language mixing, commonly referred to as "code-mixing," is typical in multilingual communities. Non-native English speakers who are multilingual often use Anglicization'sin their primary language and English-based phonetic typing to code-mix. Along with language mixing at the sentence level, code-mixing behaviour at the word level is rather common. This linguistic phenomenon poses a significant challenge to conventional NLP systems, which currently manage the integration of many languages using monolingual resources. The goal of this proposal is to raise awareness among researchers of the need forsentiment analysis in social media material that has been code-mixed.

We focus on the fusions of the third and fourth most common languages in the world, respectively, English and Hindi (Hinglish). Since Hindi words are written using English alphabets, Hinglish text is composed of both English and Hindi words written in English script. Sentiment analysis is the process of evaluating the emotions expressed in a text thatwas written by a person or user. A person may have published material on a particular topicor experience on a social networking site like Facebook, Twitter, YouTube, or IMDb, on an e-commerce website like Amazon, or on the website of a company or other organisationwhere they discussed a product or service they utilised.

These emotions can be categorised into three categories—negative, neutral, and positive—or five categories—strongly negative, negative, neutral, positive, very positive, and so on—depending on the needs and preferences of the individual. According to the users' emotions,the feelings may also be categorised as happy, angry, sad, or frustrated. They can also be categorised as will purchase or will not purchase depending on the users' level of interest.

The two main methods for extracting sentiment are machine learning and lexicon-based methods. The majority of the time, machine learning algorithms are employed to extract attitudes from documents and sentences. Among supervised machine learning algorithms,Naive Bayes, SVM, and Maximum Entropy perform best. Algorithms for machine learning are often employed. A dictionary is used in lexicon-based techniques to sentiment analysis.

Many firms will be able to better understand their customers and arrive at more informed judgements with the aid of automated systems that can recognise emotions in text. The development of sentiment analysis of code mixed language is shown in this review study.

II. LITERARY SURVEY

Mishra et al. (2018)[6] found that when doing sentiment analysis on the code-mixed datasetmade up of Hindi and English words, a maximum F1-score of 0.58 could be reached by utilising the char(2,6) gramme features of Tf-Idf vectors on the SVM. They also experimented with several Tf-Idf vectorization approach variants on SVM, MLP, Bi-LSTM, and ensemble voting classifiers, including unigrams, uni-bigrams, uni-bi-trigrams, and char(3,6) gramme. Utilizing unigrams, unibigrams, uni-bi-trigrams, and char (3,6) grammefeatures on the SVM, an F1-score of 0.57 was obtained, while an F1-score of 0.55 was obtained using glove avg features on the SVM. Voting classifier achieved an F1-score of 0.55 whereas MLP and Bi-LSTM fell short of that mark.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 1, January 2023

DOI: 10.17148/IARJSET.2023.10130

5525 test tweets and 10995 mixed Hin-Eng code tweets made up the dataset utilised in thisresearch. It wasn't the same dataset that was utilised in the experiments reported in this publication.

In the study by Patra et al. (2018)[2], it was discovered that a dataset consisting of Hindi- English as code-mixed language, where the training dataset consisted of 12936 phrases andthe test dataset consisted of 5525 sentences, could reach a maximum F1-score of 0.569. Thedataset was made available to the ICON-2017 participants in the shared task Sentiment Analysis of Indian Language (Code Mixed).Sentiment Analysis performed by Gaurav Singh (2020)[9] on the hinglish dataset was assessed using sensitivity. Support Vector Machines, KNN, Decision Tree Classifiers, Gaussian Naive Bayes, Multinomial Naive Bayes, Logistic Regression, Random Forests Classifier, and ensemble voting classifier were used in the experiment.

It was observed that the Ensemble Voting (soft) classifier achieved the best F1-score of 0.6907 for classifying the Hinglish code-mixed data from all the experiments conducted above. Out of all the studies, the logistic regression model produced the best results in 6. According to the results of the studies, it was followed by Random Forest, SVM, Multinomial Nave Bayes, KNN, Decision Tree, and Gaussian Nave Bayes in terms of ranking.

Table 1: LITERARY SURVEY

Year	Title	Author	Description	Result
2016	Sentiment Analysis of codemixed data set containing Hindi and English words	Ravi and Ravi	They used the RBFNet (Radial Basis Functional Neural Network), SVM, Decision Tree,Logistic Regression, Random Forest, MLP, and Nave Bayes algorithms to do sentiment analysis.	They discovered that the RBF Neural network performed the best.
2018	Hinglish Sentiment Analysis	Mishra et al.	10995 mixed training and testtweets in Hin-Eng code were included in the dataset utilisedfor this investigation.	With the use of the char(2,6) gramme characteristics of the Tf- Idf vectors on the SVM, amaximum F1-score of 0.58 could be attained.
2018	Hinglish dataset analysis	Patra et al.	The dataset was made available to the ICON-2017 participants inthe shared task Sentiment Analysis of Indian Language (Code Mixed).	The SVM and ensemble voting classifier were combined to provide the highest F1-score possible.
2019	Sentiment Analysis of codemixed YouTubecomments (Hinglish)	Kaur et al.	The data were converted using the vectorization techniques Tf- Idf vectorizer, count vectorizer,and term frequency vectorizer.	It was found that usingthe Tf-Idf vectorizer asthe data transformationapproach produced thebest results.
2020	Sentiment Analysis of Code-Mixed Social Media Text (Hinglish)	Gaurav Singh	The F1-score was used to assessthe developed models.	The best F1-score was obtained by Ensemble Voting (soft) classifier, which was 0.6907.

From the above research papers, we inferred:

- During the process of the project , data preprocessing is crucial as it increases the dataefficiency and decreases response time.
- To provide the best result more models should be compared .
- Random forest algorithm is very slow and requires huge amounts of calculations andas such could not work upon large data sets without vectorizers.
- We selected the hinglish data set which was used previously by Gaurav Singh[9] in2020 for better comparison and survey.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 1, January 2023

DOI: 10.17148/IARJSET.2023.10130

III. DATA PRE-PROCESSING

Text pre-processing is achieved through

- Tokenization
- Stemming
- Stop words removal

[1] **Tokenization**: Tokenization is the process of dividing text into a list of tokens from a string or text. Tokens can be viewed as components; for example, a word in a phrase isa token, and a sentence in a paragraph is a token.
It is achieved through the *nltk.tokenize* library.

[2] **Stemming**: Stemming typically refers to a rudimentary heuristic method that removesderivational affixes from the endings of words in the hopes of attaining this aim most of the time.

[3] **Stop Word Removal**: A stop word is a frequently used term that a search engine has been configured to ignore, both while indexing items for searching and when retrievingthem as the result of a search query. Examples of stop words include "the," "a," "an," or "in."

IV. DATA TRANSFORMATION

The data was transformed into vectors of numbers using several vectorization techniques usingthe predefined libraries in python. The vectorization techniques used were Count Vectorizer, One Hot Binarizer and Tf-Idf Vectorizer.

(I) Count Vectorizer

Create a token count matrix out of a group of text documents. Using scipy.sparse.csr_matrix, this implementation creates a sparse representation of the counts.

The number of features will be equal to the vocabulary size discovered by analysing the data if an a-priori dictionary is not provided and if an analyser without feature selection is not used.

(II) One Hot Binarizer

Data should be binarized (feature values should be set to 0 or 1) based on a threshold. When avalue exceeds the threshold, it maps to 1, but when it is equal to or less than the threshold, it maps to 0. Only positive values map to one when the default threshold is set to 0.

An analyst can choose to just take into account a feature's existence or absence rather than, say,a quantifiable number of occurrences when binarizing text count data.

(III) Tf-Idf Vectorizer

Create a TF-IDF feature matrix from a group of raw documents.similar to CountVectorizer, then TfidfTransformer.

V. ALGORITHMS OVERVIEW

(I) Multi-layer Perceptron classifier (MLP)

A dataset is used to train the multi-layer perceptron (MLP), a supervised learning technique, tolearn the function f():RmRo, where m is the number of input dimensions and o is the number of output dimensions. It is possible to learn a non-linear function approximator for either classification or regression given a collection of features X=x1,x2,...,xm and a target y. There may be one or more non-linear layers, known as hidden layers, between the input layer and theoutput layer, which distinguishes it from logistic regression.

from sklearn.neural_network import *MLPClassifier*

the above library is used to create the model

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 1, January 2023

DOI: 10.17148/IARJSET.2023.10130

(II) Convolutional neural networks (CNN)

A neural network with at least one convolutional layer as a layer. Layers of a convolutional neural network often include any or all of the following:

- **convolutional layers**
- **pooling layers**
- **dense layers**

In some types of issues, including image recognition, convolutional neural networks have achieved remarkable success.

VI. EVALUATION METRICS

(i) F1 Score: Calculate the balanced F-score, sometimes referred to as the F-measure or F1 score.

The F1 score may be thought of as a harmonic mean of accuracy and recall, with the highest value being 1 and the poorest being 0. Precision and recall both contribute equally in terms ofpercentage to the F1 score. The F1 score is calculated as follows:
F1 = 2 * (precision * recall) / (precision + recall)

This is the average F1 score for each class in the multi-class and multi-label situation, with weighting based on the average parameter.

VII. RESULT

On the cleansed data, feature extraction methods like Count Vectorizer, One HotBinarizer, and Tf-Idf vectorizer were employed –

(I) Count Vectorizer

(a) Multi-layer Perceptron classifier (MLP)

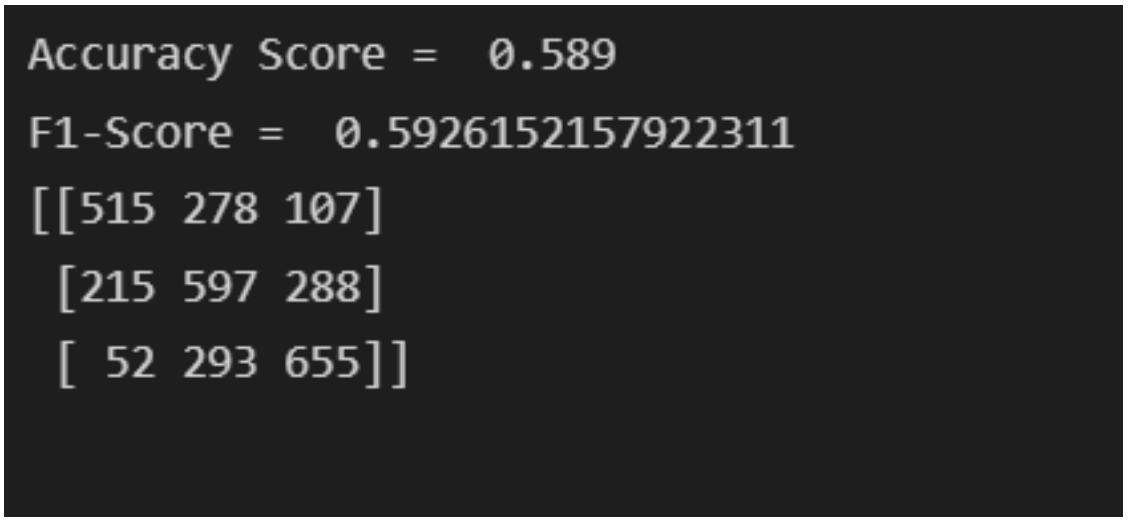

Figure 1: Score for accuracy and F1 for MLP

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 1, January 2023

DOI: 10.17148/IARJSET.2023.10130

(b) Convolutional neural networks (CNN)

```
Accuracy Score =  0.6216666666666667
F1-Score =  0.6252936784599569
[[591 229  80]
 [244 611 245]
 [ 67 270 663]]
```

Figure 2: Score for accuracy and F1 for CNN

(II) One Hot Binarizer

(a) Multi-layer Perceptron classifier (MLP)

```
Accuracy Score =  0.5843333333333334
F1-Score =  0.5881272174749389
[[538 259 103]
 [238 570 292]
 [ 63 292 645]]
```

Figure 3: Score for accuracy and F1 for MLP

(b) Convolutional neural networks (CNN)

```
Accuracy Score =  0.624
F1-Score =  0.6277235116850656
[[597 232  71]
 [258 598 244]
 [ 58 265 677]]
```

Figure 4: Score for accuracy and F1 for CNN

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 1, January 2023

DOI: 10.17148/IARJSET.2023.10130

(III) Tf-Idf Vectorizer

(a) Multi-layer Perceptron classifier (MLP)

```
Accuracy Score =  0.575
F1-Score =  0.5793089579931577
[[508 288 104]
 [217 586 297]
 [ 56 313 631]]
```

Figure 5: Score for accuracy and F1 for MLP

(b) Convolutional neural networks (CNN)

```
Accuracy Score =  0.6626666666666666
F1-Score =  0.6670798644081799
[[631 232  37]
 [263 638 199]
 [ 36 245 719]]
```

Figure 6: Score for accuracy and F1 for CNN

The above results were obtained on a system with the following specification:

- Intel ® core(™) i5-9300H CPU @ 2.40 GHz processor
- 16 GB SODIMM RAM
- Intel ® UHD Graphics
- NVIDIA GeForce MX230

VIII. CONCLUSION

Our survey has shown various score on using these models. Our surveey shows the capabilityof already available classification algorithms and comparing them for a best fit for our dataset.The Tf-Idf vectorizer was proved as the vectorization strategy to provide the best results in thestudy report by Gaurav Singh[9], however this time it was different since different vectorizersproduce different outcomes when employing various models.

Multi-layer Perceptron classifier gave the best outcome when count vectorizer was used.

Convolutional neural networks gave the best outcome when tf-ldf vectorizer was used.

Using a spell-checking dictionary to normalise Hindi words considerably improved the classifiers' performance. The

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.12 ⁜ Vol. 10, Issue 1, January 2023

DOI: 10.17148/IARJSET.2023.10130

classifiers performed worse when emoticons were converted to keywords, which may have been because people in the text sometimes used caustic emoticons.So perhaps it was unable to communicate the emotional range of the tweets.

Hence, the dataset was not iterated further.

Table 1: F1 Scores

F1-Score	MLP	CNN
Count Vectorizer	0.5926	0.6252
One Hot Binarizer	0.5881	0.6277
Tf-Idf Vectorizer	0.5793	0.667

IX. FUTURE SCOPE

Several neural network designs, including RNN, GRU, and Bi-RNN, The LSTM and BERT can be tested. In a same manner, several additional data variances Experimental transformationmethods include the Count Vectorizer and the Tf-Idf. Word2Vec embeddings, One Hot Binarizer, Doc2Vec embeddings, Glove embeddings as well as Fasttext embeddings.

REFERENCES

[1] G. I. Ahmad and J. Singla, "Sentiment Analysis of Code-Mixed Social Media Text (SA-CMSMT) in Indian-Languages," 2021 International Conference on Computing Sciences (ICCS), 2021.
[2] Patra, B., Das, D. and Das, A. 2018. Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL_Code-Mixed Shared Task @ICON-2017.
[3] Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2018) supervised learning of universal sentence representations from natural language inference data.
[4] SemEval-2020. 2020. SemEval-2020 International Workshop on Semantic Evaluation. [Online]. [Accessed 10 February 2020].
[5] Ravi, K. and Ravi, V. 2016. Sentiment classification of Hinglish text. Pp.641-645.
[6] Mishra, P., Danda, P. and Dhakras, P. 2018. Code-Mixed Sentiment AnalysisUsing Machine Learning and Neural Network Approaches.
[7] Patra, B., Das, D. and Das, A. 2018. Sentiment Analysis of Code-Mixed Indian Languages: An Overview of SAIL_Code-Mixed Shared Task @ICON-2017.
[8] Kaur, G., Kaushik, A. and Sharma, S. 2019. Cooking Is Creating Emotion:A Study on Hinglish Sentiments of YouTube Cookery Channels Using Semi-Supervised Approach. Big Data and Cognitive Computing. 3
[9] Gaurav Singh, 2020. Sentiment Analysis of Code-Mixed Social Media Text (Hinglish) using vectorization and models. Available from: https://arxiv.org/abs/2102.12149