Interim Report

Project Title: Studying the Effect of Vectorization Techniques in Mix-Code (Hinglish Language) on Open-Source Data Using Machine Learning and Transfer Learning Methodology.

Researcher Name: Murthy S Routhula

Researcher ID: D00243413

Researcher Email: d00243413@student.dkit.ie

Research Coordinator: Dr. Abhishek Kaushik

Course: M.Sc. Data Analytics

Department: Computer Science and Mathematics

Institution: Dundalk Institute of Technology

Place: Dundalk, Ireland

Year: 2021-2022

Acknowledgement

Table of Contents

Abstract:

One of the popular virtual learning sources in the present world is YouTube which has been accessed by billions of Internet users. Due to its popularity, the number of YouTubers has increased. Generally, people show their intentions about the videos posted on YouTube through comments. India has a population of 1.4 billion (India Population (2022) - Worldometer 2022) and has nearly 121 languages and 270 mother tongues (Jo Hartley 2021). Hindi is one of the most spoken languages in India. Indians mostly use Mix-Code language in commenting i.e., Hinglish which is the combination of Hindi and English languages. This project will be useful in analyzing the Mix-Code YouTube comments given by users for the videos posted by YouTubers. It helps in knowing the intention of users according to the video content and helps YouTubers to post videos with better quality and content. Different Vectorization techniques using Term Frequency – Inverse Document Frequency (TF-IDF), Term Frequency, Count Vectorizer, Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), Cross-Lingual Language Model (XLM), etc. are to be applied to the datasets to transfer comments to features. Supervised learning models both parametric and non-parametric models are planned to be trained using these vectorized datasets along with labels which include different classes like Questions, Suggestions, Gratitude, etc. This conduction of different combinations is to check the best prediction model based on the different evaluation methods for the Hinglish Mix-code.

Keywords:

Natural Language Processing, Sentimental Analysis, YouTube, Internet, Mix-Code, Hinglish, Machine Learning, Vectorization, Evaluation methods.

1. Introduction

YouTube is an online video-sharing social media platform that started on 14th February 2005 and is owned by Google on(Matthew Johnston 2022) November 13, 2006. It has billions of monthly users who watch videos for billions of hours collectively for their requirements. As it is one of the best learning and research platforms, it has expanded into mobile platforms too (William L. Hosch 2022). The videos on YouTube include short films, movies, documentaries, cooking channels, educational and technological related, etc. Everyone has their food preferences. Especially international students who have habituated to the home food learn to cook food themselves using YouTube videos. Due to this reason, many YouTubers started doing videos based on cooking different cuisines which some channels are very popular for their unique content. To know about the viewers' intentions and feedback on the videos, they must manually read the comments and prepare for the next video and improve. This will take a lot of time if comments are more than hundreds. This project can help in finding the nature of the comment user has given for the uploaded video instead of manual reading. This will be achieved by training the model with different types of comments with labels to understand the patterns and predict the new comments label.

This Project comes under Sentimental Analysis using Natural Language Processing popularly known as NLP. NLP started in the 1950s and is supported by Alan Turing's article titled "Computing Machinery and Intelligence" popularly known as "Turing Test" which automates the assumptions and generation of Natural Language (Natural Language Processing - Ela Kumar - Google Books n.d.). "*Sentimental Analysis which is also called opinion mining is Natural Language Processing technique used to determine whether the text data is positive or negative or neutral*" (Sentiment Analysis Guide 2020). These texts may be extracted from different comments, reviews, paragraphs, etc. It is mainly applied to social media, surveys, customer services, etc. In NLP as the natural language is processed which is stored in the form of documents or tables, the main words are extracted and used to get the opinion of the text. These words are converted to vectorized forms using different vectorization methods as mathematical calculations can be done on numerical data. This vectorized data will be trained to

Machine Learning (ML) model. Generally, Classification models are integrated into the Natural Language Processing processes. This is because different texts should be classified based on the nature of the text data which may be positive or negative or neutral. As labels will be provided for training the model, Supervised learning will be applied in this project.

Machine Learning (ML) is a term introduced by Arthur Samuel in 1952 while he was writing the computer program to play checkers game (A Short History of Machine Learning -- Every Manager Should Read n.d.). It involves mainly two types of learning namely Supervised and Unsupervised.

- In Supervised Learning, the Machine Learning models are trained on data called training data that consists of already assigned labels. Then the model is tested using test data to check the prediction capacity. The evaluation is conducted based on the actual test results and predicted results to check the accuracy of the models.
- In Unsupervised Learning, no labels will be provided, and the data will be clustered based on the patterns recognized in the model. In this project, the data has Mix-Code textual comments, and labels were assigned based on the type of comment, Supervised Learning models are trained with the vectorized Mix-Code text along with the labels.

Mix-Code languages consist of two or more language varieties while using. This type of language can be usually observed in general conversation, the local language, comments, reviews, etc. Hinglish is one of its types and it is a mix of Hindi and English Languages as shown in Figure 1. Red colour font words belong to Hindi language vocabulary and blue colour font words belong to English vocabulary. They are both used to form a meaningful sentence whose meaning can be seen. The data consists of most of these types of comments. There are some challenges in analyzing the Mix-Code languages as stop words in Natural Language Processing should be given manually depending on our requirements. Some of the other Mix-Code languages can be noted in Table 1.

> **Example I**
>
> HINGLISH: ye ek code mixed sentence ka example hai
> ENGLISH : this is an example code-mixed sentence
>
> **Example II**
>
> HINGLISH : kal me movie dekhne ja raha hu. How are the reviews?
> ENGLISH: I am going to watch the movie tomorrow. How are the reviews?
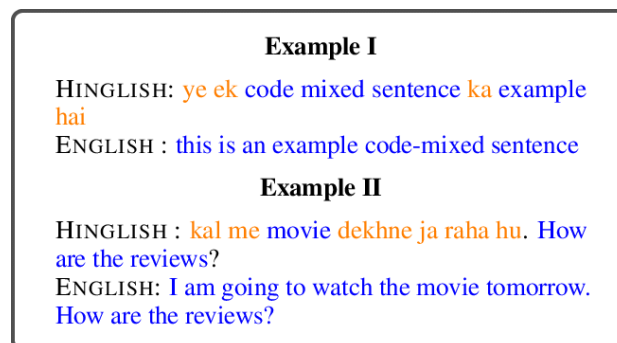
Figure. 1. Hinglish Mix-Code Language. *Source:* (Srivastava and Singh 2021)

Mix-Code	Languages
Benglish	Bengali and English
Chinglish	Chinese and English
Denglisch	Deutsch (German) and English
Dunglish	Dutch and English
Greeklish	Greek and English
Poglish	Polish and English
Porglish	Portuguese and English
Spanglish	Spanish and English
Svorsk	Swedish and Norwegian
Tanglish	Tamil and English

Table. 1. Mix-Code Language Types (Uma Gunturi 2020)

The flow of this project includes cleaning data like removing special characters, smiley symbols, etc. Different types of vectorizations are planned on the data namely TF-IDF, Term Frequency (TF), Count Vectorizer, BERT transformers, etc. Supervised learning is to be applied to all the transformed data vector forms with different classification models like Logistic Regression, K-Nearest Neighbors, Naïve Bayes, Decision Trees, Random Forests, Support Vector Machine, etc. This Report is divided into 7 sections namely Introduction, Literature Review, Methodology, Evaluation, Ethical Considerations, Conclusion and Future Work, and References. The problem statement, the structure of the report, research questions, and research motivation is discussed in the Introduction. The background research works, methods, and influenced works are mentioned in the Literature review. The methodology of how the project has been planned and detailed steps of implementation are discussed in the Methodology section. The description of data and pre-processing steps are mentioned in Data Exploration and Pre-Processing. The Ethical methods regarding the project and data are discussed in Ethical Considerations. The progress of the project and hypothesis explanation are discussed in the Conclusion and Future work section. The work references are added in the References section.

Research Questions

1. Which vectorizer techniques can be effectively used for Machine Learning models on Hinglish Mix-Code?
2. Which parametric or non-parametric model is the best performing model on Hinglish data?
3. Is Principal Component Analysis (PCA) and Independent Component Analysis (ICA) on the Machine Learning models help in getting good results for Mix-Code models?

2. Literature Review

This section briefly discusses the literature survey and background studies done for this sentimental analysis.

Data Pre-processing

Data pre-processing includes data cleaning, feature extraction, etc. Data cleaning consists of the removal of stop words, line breaks, emojis, etc. The feature extraction methods used in this analysis are count vectorizer, TF-IDF, term frequency, and transformers like BERT, GPT, and XLM. Kumar et al. used a TF-IDF vectorizer to extract features from Amazon's electronic items dataset and input them into the SVM algorithm (Kumar and Subba 2020). Irawaty et al. made the vectorizations comparison to analyze YouTube comments on Nokia products. They have used TF-IDF, Count vectorizer, and hashing vectorizer for vectorization. They have used K-Nearest Neighbor, SVM to classify. Their evaluation results show that TFIDF with SVM has good accuracy of 97.5% than other combinations (Irawaty et al. 2020). Shah et al. have conducted a Sentimental Analysis of Marglish comments on YouTube cookery channels. Marglish is the mixed code of Marathi and English. They have achieved the best accuracy of 62.68% for the combination of the Count Vectorizer and Multilayer Perceptron. The best models they suggested for Marglish datasets are Multilayer Perceptron and Bernoulli Naïve Bayes (Shah et al. 2020). Aro et al. analyzed the effect of removing stopwords on text data classification of SMS spam datasets. They have modeled using a Decision tree and Multinomial Naïve Bayes. They have found that the removal of stopwords has no effect on the classification effect of text mining but reduced the confidence level of prediction (Aro et al. 2019). AbdulNabi et al. used deep learning for spam mail detection. They have used BERT (Bidirectional Encoder Representations from Transformers) transformer which was pre-trained and fine-tuned to separate spam mails from non-spam. Then they were trained and tested by Machine Learning algorithms using two separate datasets (AbdulNabi and Yaseen 2021). Devika et al. worked on extracting the key phrases from social data using the sentence transformer of the BERT model. As BERT can enhance the performance in Natural Language Processing tasks and extract typical phrases in tweets, their model of BERT with sentence transformer gave an accuracy of 86% which is higher than their other models (Devika et al. 2021). Qu

et al. did the emotion classification of Spanish language data with XLM-RoBERTa for word embedding and the transformer encoder for feature extraction. The extracted features are given to the TextCNN model as inputs (Qu et al. 2021). Kadriu et al. used a Bag of words and word analogies for Albanian text classification. The text has been classified using two approaches, one is converting the text into vector space and the second is using FastText for hierarchical classification. For classification, the bag of words model gave the best evaluation result. For multi-label text, FastText gave better performance. Overall, using the bag of words model gave 94% of accuracy (Kadriu et al. 2019).

Mix-Codes

Mix codes are combinations of different languages in conversations. The data used in this analysis consist of Hinglish mix code which is a combination of Hindi and English. This language is mostly used in India in casual conversations and commenting on social networks. Agarwal et al. worked on Hinglish dialogue generation. They have used mBART multilingual sequence-to-sequence transformers for Hinglish dialog generation which sets new benchmarks for mix codes dialog generation tasks (Agarwal et al. 2021). Kumar et al. used neural networks and transfer learning for cyberbullying detection on mixed code data. They have included typography learned using Machine Learning Processing along with English and Hindi languages. They have combined those features to the unified level which gives the unique distribution advantage without increasing the input space dimensionality (Kumar and Sachdeva 2020). Mundra et al. evaluated text representation methods to detect cyber harmful content on social media. The data considered for analysis is in the Hindi and English mix-code popularly known as Hinglish. In their analysis, it is found that character-based embedding is working well for noisy data. This model also worked better than pre-trained word embedding (Mundra and Mittal 2021). Singh et al. conducted a Sentiment Analysis on social media mix-code content which is in the Hindi and Punjabi languages. The labels of the data include positive, negative, and neutral based on the words in the text. They have used the N-gram approach applied to the sentence (Singh and Goyal 2020). Bansal et al. experimented with Sentiment Analysis on English Punjabi mix-code social media data. They have collected data through Twitter and Facebook APIs. They have used a pipeline Dictionary vectorizer and an N-gram approach (Bansal et al. 2020).

Machine Learning

Machine Learning is the branch of Artificial Intelligence that is helpful in predictions on data. Both Supervised and Unsupervised Learning are useful in sentimental analysis. Unsupervised is used to cluster or separate the data based on patterns while Supervised Learning is used to train the model based on the outputs which help in future prediction. Bhavitha et al. applied Machine Learning algorithms to subjective data to get the intention behind the text whether it is positive, negative, or neutral regarding the newly launched product. They have got 85% of accuracy on supervised learning techniques than unsupervised learning techniques (Bhavitha et al. 2017). Agrawal et al. evaluated supervised and unsupervised learning techniques in sentimental analysis. They have evaluated based on the accuracy, benefits, and disadvantages of every mechanism. They have got good metrics for supervised models when compared to unsupervised models (Agrawal et al. 2021). Bansal et al. experimented with Sentiment Analysis on English Punjabi mix-code social media data. Machine Learning models used are Decision tree, Gaussian Naïve Bayes, and Logistic Regression. The evaluation metrics of Logistic Regression are better with an accuracy of 86.63% and an F1 score of 88% when compared with other models (Bansal et al. 2020). Harfoushi et al. analyzed Twitter data which consists of opinions of individuals, images, and tweets. They have implemented Azure Machine Learning models like SVM and Logistic regression. The results confirmed that Microsoft Azure Algorithms can be used to build effective models when compared to the traditional way of modeling in data analytics (Harfoushi et al. 2018). Thelwall M has checked if there is an effect of gender bias in Machine Learning for Sentiment Analysis. He has trained and tested the models using three sets of datasets of hotel and restaurant reviews. His study declares that mixed gender datasets are preferring the opinion of women.

Conclusions are that the training of the model on the same gender improves the performance of the model less than adding additional data on both genders' data (Thelwall 2018). Valencia et al. predicted the price movement of cryptocurrencies using Machine Learning and Sentiment Analysis. Models like Neural Networks, Support Vector Machines and Random Forest have been implemented based on the data from Twitter and the market to analyze the price movement of Bitcoin, Ripple, Ethereum, and Litecoin. Results indicate that using Machine Learning price prediction can be possible and Neural networks are better in performance than other models (Valencia et al. 2019). Swaminathan et al. modeled hate speech identification based on the Dravidian mix-code. They have used Machine Learning, deep learning, and ensemble models. For sentiment classification, they have trained and tested the models like Naïve Bayes, Decision tree, Random Forest, Long Short-Term Memory, and AdaBoost. For Hate speech and offense content identification, they have used the models Naïve Bayes, Decision tree, Random Forest, Long Short-Term Memory, SVM, and Gated Recurrent Unit. The F1 scores obtained for Naïve Bayes and Long Short-Term Memory are 61% and 60% respectively. For hate speech identification, subtask A of LSTM gave an F1 score of 50.02% and subtask B of the ensemble approach gave an F1 score of 24.26% (Swaminathan et al. 2020).

Sentiment Analysis

Sentiment Analysis is the process of extracting the intentions from the text computationally along with identifying and categorizing the opinions. It is a sub-field of Natural Language processing to get the positive or negative or neutral opinion of the text. Fang et al. conducted sentimental analysis on online product review data from Amazon.com. They have analyzed sentence-level categorization and review-level categorization (Fang and Zhan 2015). Serrano-Guerrero et al. worked on a comparative analysis of some free web services. They analyzed using the reviews based on three different collections and analyzed each tool (Serrano-Guerrero et al. 2015). Williams et al. investigated the effect of idioms in Sentiment Analysis. They evaluated models based on precision, recall, and F1-score. The statistical significance of improvement was confirmed using McNemar's test (Williams et al. 2015). Nguyen et al. built a model for analyzing stock movement using sentiments from social media. They have achieved better accuracy while analyzing the 18 stocks using one-year transactions than the historical price method and human sentiment method (Nguyen et al. 2015). Alsaffar et al. performed Sentiment Analysis on the Malay language using K-Nearest Neighbor. They have used Lexicon based approach which derives the intention from text based on the words' semantic orientation. Their hybrid method outperforms the of-the-art unigram baseline method (Alsaffar and Omar 2015).

3. Methodology

In this section, the methods and flow of sentimental analysis that will be conducted are discussed. The flow of the project is divided into different sections as below.

1. Data Collection: The data is collected from the UCI website (UCI Machine Learning Repository: Youtube cookery channels viewers comments in Hinglish Data Set n.d.). The data contains the comments received by the two YouTube cookery channels namely, Nisha Madhulika's Cooking channel and Kabita's Kitchen. The data consists of labels divided into 7 categories as shown in Table 2.
2. Data Preprocessing: The raw data consists of many line breaks and smiley symbols. They will be removed in the preprocessing stage.
3. Data Visualization: The Visualization Analysis will be carried out to analyze labels, stop words, hashtags, word counts, character counts, numerical values present, etc.
4. Vectorization: The processed data will be converted to vector form datasets using different vectorization techniques like Term Frequency-Inverse Document Frequency (TF-IDF), Term Frequency (TF), Count Vectorizer, BERT Transformers, etc.

5. Feature Scaling: Different Scaling techniques will be applied to check the effect of scaling on the Machine Learning evaluation results.
6. Machine Learning: The Machine Learning models are trained and tested with the vectorized datasets. Different cross-validation techniques will be used for each model. The training data will be 70% and the testing data will be 30%. The dimension reduction technique like Principal Component Analysis and Information separation technique like Independent Component Analysis will be performed.
7. Evaluation: As the Sentimental analysis is based on the classification type of supervised learning, the evaluation will be done based on Precision, Recall, F1 Score, Confusion matrix, Classification report, Accuracy, Area Under Curve, etc.
8. Results: The best results for the research question will be fixed based on the evaluation results of the different Machine Learning models applied to different vectorized datasets.
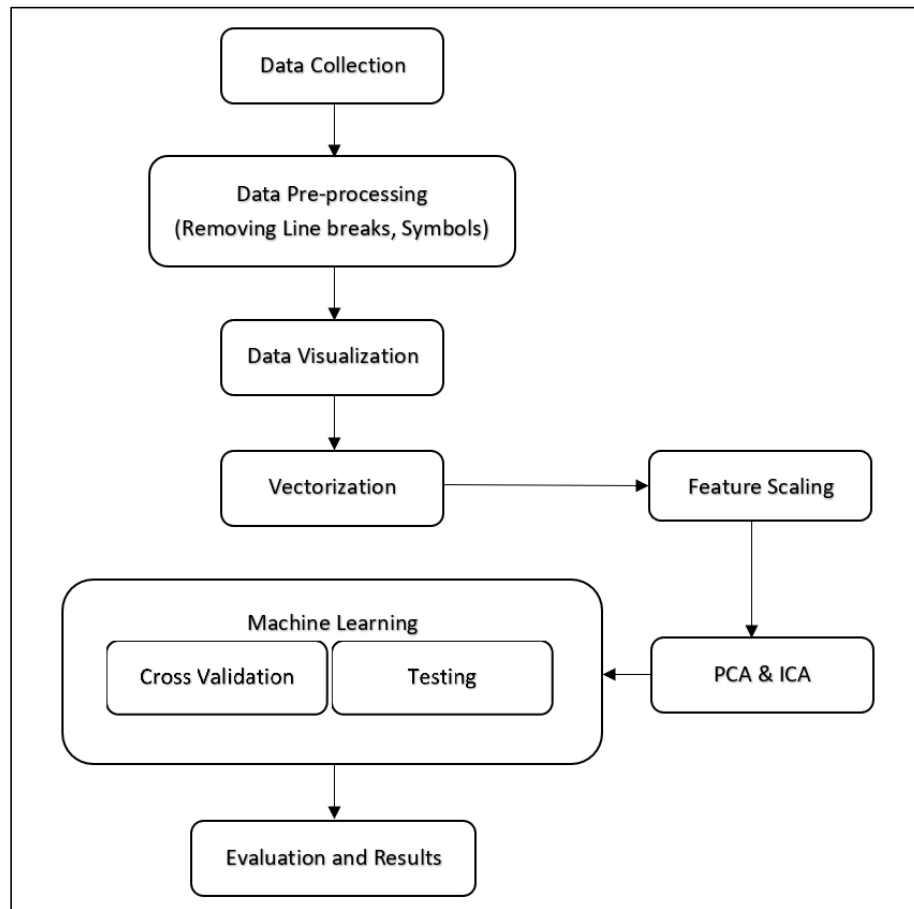

Figure. 2. Flow of Methodology

3.1 Data Collection

The two datasets are of two YouTube Cookery channels taken from the UCI website. The channels are India's popular cooking channels namely NishaMadhulika and Kabita's Kitchen. Each dataset consists of 4900 rows. Each row has a comment given by the user and the type of user intention through the comment. The comments were clustered and labeled using the unsupervised learning method Density-Based Spatial Clustering of Applications with Noise (DBSCAN) after collecting the YouTube comments through its API in March 2019 (Kaur et al. 2019).

The dataset labels were classified into 7 categories based on the viewers' intentions. Those 7 categories include Gratitude, About Recipe, About Video, Praising, Hybrid, Undefined, Suggestion, or Query. The

description of each label can be seen in Table 2. The number of rows of each dataset was divided equally according to those 7 labels as shown in Table 3.

Label Class	Label Type	Label Description
1	Gratitude	This Label indicates that the comment is the gratitude shown by the viewer to the YouTuber. Examples: 1. Thank you so much for putting this detailed video 2. thank u mam 3. thank you didi
2	About Recipe	This Label indicates that the comment is the review given by the viewer about the recipe how good it is and tastes. Examples: 1. This is a perfect biryani recipe 2. Nice recipe, that was so simple yet delicious 3. 2 good Mam very nice recipe
3	About Video	This Label indicates that the comment is the review given by the viewer about the video how good it is and playtime. Examples: 1. AMAZING! Maine ye video dekhkar dum biryani banana sikha hai 2. very nice video mam, Great video! 3. nice video
4	Praising	This Label indicates that the comment is the review given by the viewer praising the chef and admiring him. Examples: 1. the way u cook, it's really looking so beautiful 2. Very nice cooking style 3. Super your recipes are amazing
5	Hybrid	This Label indicates that the comment includes two or more qualities of labels. For example, the viewer expresses his views about the recipe and video in the same comment. Examples: 1. Thakuuu soo mch mam u r such a talented

		2. Nice Aunty ji..........kaun se oil ka use karna hoga??
		3. hello nisha,ive tried ur alo paratha n it was just awesome,i just love u n ofcourse ur recipes.
6	Undefined	This Label indicates that the comment doesn't come under any of the other labels like praising or showing gratitude or querying about recipes or videos. Examples: 1. I am hungry 2. Who try this please one like 3. Happy new year aanti
7	Suggestion or Query	This Label indicates that the comment is the question or suggestion by the viewer about the recipe. Examples: 1. Atta flour means wheat flour? 2. Can we grate the potatoes mam? 3. Kya stafing me Magi masala dal sakte he

Table. 2. Labels indication for the comment type and description

Labels	Nisha Madhulika Dataset	Kabita's Kitchen Dataset
Label-1	700	700
Label-2	700	700
Label-3	700	700
Label-4	700	700
Label-5	700	700
Label-6	700	700
Label-7	700	700
Total Comments	4900	4900

Table. 3. Distribution of Labels in the Datasets

3.2 Data Preprocessing

YouTube comments given by users consist of many spelling mistakes and special characters. This is because the comments resemble the common conversation type language. To make the data efficient for modeling, preprocessing will be done on both datasets. Pre-processing includes the removal of special characters, smiley symbols, numbers, line breaks, converting text to lowercase, stop words, etc. Tokenization will be done before vectorization.

Special characters include punctuation marks. Smiley symbols are generally used on social media to replicate the expressions. So, they will be removed. Line breaks occur if the user tries to write 2 different reviews in the same comment. All the text will be converted to lowercase to attain equality in the strings

while performing the vectorization. Stop words are the most used words in sentences. For example, stop words are like 'at', 'is', 'was', 'if', etc. But these stop words should be configured according to the use case. As the comments used for analysis are of Hinglish mix-code language, we should manually add stop words according to our requirements. Tokenization means the splitting of sentences into keywords, phrases, etc called Tokens by removing spaces, punctuations, etc.

3.3 Data Visualization

The main purpose of this data visualization is to analyze the data and understand it more clearly. It provides a well-organized visual representation of data to easily analyze and interpret the understanding. The distribution of labels, stop words, hashtags, word counts, character counts, numerical values present, etc in the data will be analyzed using visualizations. This will be achieved by plotting the graphs like Boxplots, Count plots, etc. using matplotlib or seaborn libraries.

3.4 Vectorization

In Machine Learning, while working with categorical data, we need to convert them to numerical as the statistical calculation can be done only on numerical values. For this requirement, there are numerous methods to convert categorical data into numerical data. Some of the methods are dummies creation, Values assignment, Vectorization, etc. In vectorization, the text is tokenized and converted into vectors called Feature Extractions. One of the best methods for this feature extraction is Bag of Words. In the Bag of Words model, the grammar and order of words won't be considered instead it will keep the count of word repetition. The Example of the Bag of words application can be seen in Table 4. As Bag of words feature extraction is best for classification models, this method of feature extraction will be applied before modeling.

Normal text	This Project is based on Natural Language Processing. Natural Language Processing is formerly called NLP.

Bag of Words model	BoW1 = {
	    "This":1, "Project":1, "is":2, "based":1, "on":1, "Natural":2,
	    "Language":2, "Processing":2, "formerly":1, "called":1,
	    "NLP":1
	    }

Table. 4. Bag of Words Example

The Bag of Word models used for the analysis is Term Frequency – Inverse Document Frequency (TF-IDF) Vectorizer, Term Frequency (TF) Vectorizer, and Count Vectorizer.

3.4.1 Term Frequency – Inverse Document Frequency Vectorizer: The approach in this method is that the words that are more common in one text and less common in other texts should be given high weights. For this method also, the first step will be tokenization. TF-IDF value of each word in the text will be calculated.

TF value can be calculated by,

$$TF = \frac{Frequency\ of\ the\ word\ in\ the\ sentence}{Total\ number\ of\ words\ in\ the\ sentence}$$

IDF value can be calculated by,

$$IDF = \log\left(\frac{Total\ number\ of\ sentences\ (documents)}{Number\ of\ sentences\ (documents)\ containing\ the\ word}\right)$$

$$TF - IDF = TF * IDF$$

TF value of word changes from document to document but IDF value of word remains constant as it depends on the total number of documents

3.4.2 Term Frequency Vectorizer: It is the value of TF from the TF-IDF vector without IDF value. The Term frequency of words will be calculated by dividing the frequency of words

in the sentence by the total number of words. The value of the word which is repeated more will be given preference.

3.4.3 Count Vectorizer: It calculates the value by one-hot encoding which means the value depends on the number of times the word repeats in the text. For every occurrence of the word in the text, the value will be incremented by 1. If the word is not present in the feature, it will be added. The example of count vectorization is explained in Table 5.

Normal text		Hi, how are you? Are you fine?				
Count Vectorization	Indexing	are	fine	hi	how	you
		0	1	2	3	4
	Vector values	2	1	1	1	2

Table. 5. Count vectorization

Along with the vectorizers, word embeddings of transformers like BERT, GPT, and XLM are used to convert the comments to vector formats. Word embeddings mean converting words to vectors in lower-dimensional space. By this, we can use mathematical operations on the numerical form of words in Machine Learning. Transformers are the deep learning encoder-decoder model which uses the self-attention mechanism weighting the parts of input data. They are increasing their choice for Natural Language Processing replacing other deep learning models like Recurrent Neural networks (RNN), Long Short-Term Memory (LSTM), etc.

i. BERT Model: It is a Bidirectional Encoder Representation from Transformers. It is a pretrained transformer model well-suitable for Natural Language Processing. Here it is used to extract high-quality features from text data and use them for classification analysis. It has an advantage over the Word2Vec models because it captures the differences like polysemy and context. For example, the word "bank" in "robbing the bank" and "fishing by the bank" has two different word embeddings in the BERT model when compared to Word2Vec models.

ii. GPT Model: It is a Generative Pre-trained Transformer model by OpenAI. It performs Natural Language Processing tasks like answering questions, and the relation between text fragments, etc. Using generative pre-training, the model improves the understanding of language. GPT used in this project is used for word embedding in a 768-dimensional state.

iii. XLM Model: It is Cross-Lingual Language Model. It is a pre-trained transformer for the objectives like casual language modeling, masked language modeling, and translation language modeling. XLM is used for word embedding for this project.

3.5 Feature Scaling

In raw data, the values will range widely which will make Machine Learning algorithms work abnormally. So, scaling of data is needed to normalize the features of the data. Scaling of the data should be normally done in pre-processing steps of modeling. There are different types of scaling techniques like Min-Max Scaling, Standard Scaling, Normalize Scaling, Binary Scaling, etc.

3.5.1 Min-Max Scaling: It shrinks the data to the given range of values without losing the shape of the original distribution. By default, it will scale the data in the range of 0 to 1. The scaling of data between the required range of values (a,b) is generally done by the below formula.

$$x^1(Scaled) = \frac{(b-a)(x - x_{min})}{x_{max} - x_{min}} + a$$

3.5.2 Standard Scaling: The Standard distribution is mainly achieved by standard scaling. The scaled value is the result of the difference between the actual value and the mean value of the feature divided by the standard deviation of the feature.

$$x^1(Scaled) = \frac{x - \bar{x}}{\mu}$$

3.5.3   Normalize Scaling: Normalizer is mainly used to control the size of vector to avoid numerical instabilities due to outliers. It shrinks the data between 0 to 1. It is mostly useful for regression than classification.

$$x^1(Normalized) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3.5.4   Binary Scaling: It is the technique of scaling where the threshold should be provided. The values less than or equal to the threshold will be changed to 0 and values greater than the threshold will be changed to 1. The default threshold for Binarizer is 0.

3.6 Machine Learning

Machine Learning is a branch of Artificial Intelligence where the predictions are made for future data by the algorithms based on the patterns of the data we feed while training. Machine Learning algorithms are divided into 4 types based on the data of prediction.

    a. Supervised learning: In this type, the models are trained with both Inputs and desired outputs of the data. The training data will be in the form of a matrix with the desired output in vector form called labels. One label might be the output of multiple input types. Supervised learning is further divided into Regression and Classification. In regression, the output labels are numerical data types and in classification, output labels are Categorical data types. The algorithm keeps on improving the accuracy and predictions over time based on the data.

    b. Unsupervised Learning: These models are used if the data consists of no labels to predict the output. The main purpose of this learning is to group or cluster the data based on the patterns and similarities recognized by the algorithm. Unsupervised learning is further divided into 2 types namely Clustering and Association rules. K-Means, Hierarchical, etc are important clustering types. Association rules help to find the relations and co-occurrences between features in data.

    c. Semi-Supervised Learning: It involves both unsupervised and Supervised learning models. The data which consists of no labels are clustered and provided labels using unsupervised learning. Now the data is mapped with labels and trained using supervised learning models to predict unknown future data. Based on the accuracy, the supervised learning model is again trained along with the test data.

    d. Reinforcement Learning: In this type, the model will depend on the sequence of decisions while training. The goal is to reduce the error and increase the success accuracy based on the error scenarios. The model always tries to learn from the random trails themselves.

The data for the sentimental analysis has already been labeled. So, Supervised learning models will be applied to predict the user's intention through his comment. The labels of the data should be considered as categorical as they are assigned to the sentiment types. The classification algorithms will be modeled according to this analysis's response variable data type. Based on the parameters, the supervised classification algorithms are divided into 2 types i.e., parametric, and non-parametric. Parametric models require fixed parameters and are not flexible. In non-parametric models, the parameters are not fixed. Due to this, the features increase with training data. The various parametric and non-parametric models are mentioned in Table 6. In this use case, both parametric and non-parametric algorithms are used.

Parametric models	Logistic Regression
	Bernoulli Naïve Bayes
	Gaussian Naïve Bayes
	Multinomial Naïve Bayes
Non-parametric models	Decision tree
	Random forest
	K-Nearest Neighbors
	Support Vector Machines

Table. 6. Parametric and Non-parametric models

Testing of the data will be done after modeling and training the data using parametric and non-parametric models. For testing, different cross-validation methods will be performed. Different cross-validation techniques like Test-train split, Random test-train split, k-fold, leave one out, etc will be performed to check the accuracies for different models. The data used for the testing is planned to be 30%. Based on the test results the overfitting and underfitting of models will be evaluated.

The dimensional reduction techniques like Principal Component Analysis (PCA) and Information separation techniques like Independent Component Analysis (ICA) will be applied to observe their effect on the prediction accuracy. PCA is used to reduce the dimensions of the data without losing the information. It is used to find the features that are applicable for maximum variance in the data. All the features obtained after applying PCA are orthogonal to each other. Generally, ICA will be preferred to do after PCA. ICA is used to separate information to be maximally independent. ICA is used to find the hidden factors in the features. The assumptions for applying ICA should be variables are non-gaussian and independent.

3.7 Evaluation

The evaluation metrics considered for the sentimental analysis based on classification are Accuracy, Precision, Recall, F1 Score, Classification Report, Confusion Matrix, and Area under Curve (AUC). All these metrics will be derived for all the combinations of Vectorizations, Scaling techniques, Algorithms, and Cross-validations.

a. Accuracy: It is the metric that calculates how accurately the algorithm classifies the points correctly. In classification accuracy will be calculated on True Positives, True Negatives, False Positives, and False Negatives.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Neagatives}$$

b. Precision: It is one of the model performance indicators of the classification models. The positive prediction of the model is evaluated by this metric. It is calculated by True positives and False positives predicted by the model.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

c. Recall: It is the number of true positives found by the model. It is calculated by using True positives and False negatives.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

d. F1 Score: It is used to calculate the test accuracy of the model. It is calculated using the Precision and Recall of the model by taking the harmonic mean of them. Its highest possible value is 1.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

e. Classification Report: It is one of the performance evaluation metrics that include the model's Precision, Recall, F1 score, and Support. Support is the number of actual class occurrences in the dataset

f. Confusion Matrix: It is the metric used to evaluate the predictions done by the model. The True positives, False positives, True Negatives, and False Negatives can be derived from this matrix. The number of rows and columns of the matrix depends on the number of classes in the response variable.

		Predicted class	
		Yes	No
Actual class	Yes	True Positive	False Negative
	No	False Positive	True Negative

Table. 7. Confusion Matrix Table of Classification in Supervised Learning.

g. Area Under Curve (AUC): It measures the ability of the classifier to differentiate between classes. Specificity and Sensitivity are used in finding the AUC curve. True Negative rate is called Specificity. True positive rate is called Sensitivity. Higher AUC indicates that the model is better at distinguishing the classes.

4. Results

The results obtained after modeling and testing will be compared between different parametric, and non-parametric models based on cross-validations, scaling techniques, dimensional reduction, and Information separation techniques. The results are justified based on different evaluation methods for all the combinations of techniques and models of supervised learning classification. As per the practical evaluation results and theoretical concepts from section 3, the best model will be considered.

5. Ethical Considerations

5.1 Harms and Benefits of the project

Like any other Technology and invention, Natural Language processing also has benefits and harms based on the projects it is being implemented on. As per the Confusion matrix of ethics, this research is ethically implemented through good methods yielding good results. The Project which is implemented by the Analysing sentiments on YouTube comments has more benefits when compared to the disadvantages as it saves a lot of time and manual tasks. Previously and in some present channels, the comments on YouTube are being examined for knowing the review emotion given by subscribers or viewers through manual reading and commenting. But Natural Language Processing helps in predicting the type of comment that the viewer has given, and the advanced model helps in giving a reply to the comment based on the emotion.

Figure. 3. Sentimental Analysis of YouTube comments. *Source:* (Ripul Agarwal 2020)
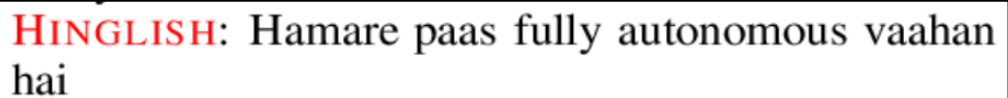
The Ethical challenges of this project include the applications that can use the implemented models on data which is either for the good or bad purpose of sentiment mining. The manual comment reviewers of social media channels or streamers will have their roles at risk as this model can replace their positions as it will predict the comment emotion in less time instead, they can upgrade themselves by developing advanced techniques for this model.

5.2 Harms and Benefits linked to the data
5.2.1 Benefits
Helps in understanding the different vectorization techniques which remove the meaning of the words for analysis by the model and instead, convert strings to numerical forms to make the model understand the patterns in data. Making an understanding of different models and the evaluation results on the data taken helps in preferring models when more data is added for training instead of starting from the initial stage. Analyzation on mix codes like Hinglish (Hindi + English), Marglish (Marathi + English), Tenglish (Telugu + English), etc. which are realistic in conversations, speech recognition systems, etc. by using this type of Natural Language processing model.

HINGLISH: Hamare paas fully autonomous vaahan hai

Figure. 4. Hinglish Text. *Source:* (Vivek Srivastava 2021)

5.2.2 Harms
Harms include the risk of not utilizing the data with consistency. The present data is transparent and consistent as the type of comments concerning emotions is equal in number. It also includes the reusability of data in the future as data used once can't be used again for training the model. There will be no concern regarding privacy for input data as the data is open-source and security policies are strictly followed while storing the data that including models, code snippets, procedures, etc.

5.3 Ethical Challenges with Dissertation
5.3.1 Collection of Data and Usage
Data is collected from UCI Machine Learning Repository. Data is open-sourced and is cited in the references as working on the same. Data is collected from Hinglish comments on two famous YouTube channels of Nisha Madhulika and Kabita. Data used for the Natural Language Processing project consists of Questions, Praise, Suggestions, Gratitude, About the Recipes, and Videos. It doesn't include any human personal information but only the comments of the people for the work mentioned on YouTube's channels. Data is not shared or sold to any third parties nor republished as it is only for study and research purposes. As the data is open source, no consent forms or privacy policies are attached to it. In case of any further research continued with the data produced by our models, they should cite this paper in their research work. The terms of our data policy will be clear and understandable indirect way, and they will be mentioned along with the report instead of clicking-through or buttons response. We can't make any changes to the data on the open-source, and we have the chance to modify it according to our use case.

5.3.2 Data Storage, Security, and Stewardship
Data is stored in such a way that its copy is available along with the if it got missed from a system. The method of remote storage will be mentioned in the report. Version control is applied to the data of research to track the changes and to revert if necessary. On systems, data is protected with a password locker. The further plan is to implement data anonymization and make the data encrypted so that even if a data breach happens, data can

be read. There is no risk in data storage for a long period as our case study doesn't include any personal information. The data is thoroughly examined and updated according to the requirement of the models in the research. Permissions for data including modifications, deletions, etc. are not given to any other people as the project is single-handled.

5.3.3 Data Hygiene and Relevance

The data collected is semi-structured and consists of sentences to analyze the type of comment. The datasets include two files of two YouTube channel comments of 4900 rows each. They have undergone different vectorization techniques after data wrangling. Different data is extracted from it like Hashtags, Numbers, Average words in the sentence, etc. to analyze the data using some visualizations. The sentences are converted to word vectors which will be Numerical data and Models are built on those vectorized data for predictions. Models include Parametric and Non-parametric Algorithms and Cross-validation is applied to the data for all Algorithms to check the correct accuracy and finalize the model for prediction. Data of models and evaluation results don't contain any sensitive information like passwords, Access codes, etc. When coming to data integrity, it will be consistent in all systems irrespective of the platform. The bias of data concerning gender or race won't be applicable here as it includes the type of comments, but no independent variable includes the nominal data. The labels of the response variable are 7 unique types and equally distributed with 700 rows each for each dataset. There will be no expiry for data in the Natural Language Processing as the synonyms of sentences don't change over time and more data can be added in the future to increase the training of the model.

5.3.4 Identifying and Addressing Harmful Bias

As all labels are considered and distributed equally in the response variable, there will be no bias in the models reducing underfitting increases the prediction level of unknown data. The non-bias nature of the datasets resembles consisting of 4900 rows each containing 7 types of 700 labels equally distributed.

5.3.5 Validation and Testing of Data Models

The challenges facing this include the finding of vectorization methods for mixed codes as mixed codes can be noticed only during normal conversation or giving comments in the native language. The stop words set for Hindi and English can't be found according to our use case so created stop words data manually based on labels taken in the target variable. To make the model more perfect in Natural language processing, the model should be continuously trained with more data for increasing accuracy in predictions. Different Algorithms are to be used for different vectorization techniques to finalize the Algorithm used for final training using cross-validation techniques.

5.4 SWOT Analysis

SWOT analysis is used to get aware of the factors while making decisions and strategies implemented (Stephen J. Bigelow 2022), it is applied to this research model for analyzing its Strengths, Weaknesses, Opportunities, and Threats. The SWOT points according to the thesis are mentioned in Table 8.

	Positive	Negative
Internal	**Strengths**	**Weaknesses**
	• Unstructured text data can be vectorized and analyzed (Amanda Porter 2022). • Accurate than human analysis. • Better understanding of market and customer satisfaction (Rachel Wolff 2020). • Saves money and time (Shemmy Majewski 2020).	• Training takes a lot of time. • Difficult to get 100% accuracy (Ximena Bolaños 2020). • Ambiguity in phrases, Words with different contexts have different meanings. • Low resource languages and mixed codes stop words need to be introduced manually (Inés Roldós 2020)
External	**Opportunities**	**Threats**
	• Application of NLP in Education (Burstein 2009). • Predictive texts, Search results, Email filters, etc. (Natural Language Processing (NLP) Examples | Tableau n.d.) • Comments Analysis, Social Media monitoring, Recruitment, etc. (Abhishek Sharma 2020) • Intelligence gathering on financial stocks and marketing research, Report Auto-generation (Ilia Lorin 2020).	• Ambiguous and vague models as they can't recognize the meaning and are unclear (Pamela Fox 2018). • Biasness of Human speech is getting stored in the machines where they show the same nature. • Loss of manual task jobs due to automated NLP applications.

Table. 8. SWOT Analysis on Natural Language Processing in Sentimental Analysis

6. Conclusions and Future Work

YouTube is one of the popular mediums for learning and gaining knowledge about new things. It also acts as an entertainment network apart from the learnings. Many videos will be uploaded on YouTube on daily basis. Many people as a part of their daily activity, like to try and learn new cooking recipes and new cuisines. Due to this, YouTubers need to focus on the quality of the content based on the users' requirements and reviews. This use case helps the cooking channel admins in adding the content supported by the users in the videos. The main aim of this sentimental analysis is to find the best combination of vectorizers, scaling techniques, and Machine Learning models on the user comments. Based on the evaluation metrics it will be decided.

The future work for this analysis includes the implementation of deep learning and neural network models on the same datasets and evaluating them for the best model. Analysis should include animations and emojis in future work. Other channel types like educational, music, sci-fi, etc topics will be covered for the sentimental analysis.

7. References

A Short History of Machine Learning -- Every Manager Should Read. Available from: https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/?sh=eb5887015e78 [accessed 26 May 2022].

AbdulNabi, I. and Yaseen, Q. (2021). Spam email detection using deep learning techniques. In: *Procedia Computer Science*. Elsevier B.V., pp.853–858.

Abhishek Sharma. (2020). *Applications Of Natural Language Processing (NLP)* [online]. Available from: https://www.analyticsvidhya.com/blog/2020/07/top-10-applications-of-natural-language-processing-nlp/ [accessed 11 April 2022].

Agarwal, V., Pooja Rao, S.B. and Jayagopi, D.B. (2021). Towards Code-Mixed Hinglish Dialogue Generation. In: *International Conference Recent Advances in Natural Language Processing, RANLP*. Incoma Ltd, pp.7–15.

Agrawal, S.C., Singh, S. and Gupta, S. (2021). Evaluation of Machine Learning Techniques in Sentimental Analysis. In: *2021 5th International Conference on Information Systems and Computer Networks, ISCON 2021*. Institute of Electrical and Electronics Engineers Inc.

Alsaffar, A. and Omar, N. (2015). Integrating a Lexicon based approach and K nearest neighbour for Malay sentiment analysis. *Journal of Computer Science*, 11(4), pp.639–644.

Amanda Porter. (2022). *What are the advantages of Natural Language Processing in AI? - Capacity* [online]. Available from: https://capacity.com/enterprise-ai/faqs/what-are-the-advantages-of-natural-language-processing-nlp/ [accessed 11 April 2022].

Aro, T.O., Dada, F., Oluwagbemiga Balogun, A. and Oluwasogo, S.A. (2019). Stop Words Removal on Textual Data Classification. *International Journal of Information Processing and Communication (IJIPC*, 7(1), pp.1–9.

Bansal, N., Goyal, V. and Rani, S. (2020). Experimenting Language Identification for Sentiment Analysis of English Punjabi Code Mixed Social Media Text. *International Journal of E-Adoption*, 12(1), pp.52–62.

Bhavitha, B.K., Rodrigues, A.P. and Chiplunkar, N.N. (2017). Comparative study of machine learning techniques in sentimental analysis. In: *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2017*. Institute of Electrical and Electronics Engineers Inc., pp.216–221.

Burstein, J. (2009). Opportunities for natural language processing research in education. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Devika, R., Vairavasundaram, S., Mahenthar, C.S.J., Varadarajan, V. and Kotecha, K. (2021). A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data. *IEEE Access*, 9, pp.165252–165261.

Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2(1).

Harfoushi, O., Hasan, D. and Obiedat, R. (2018). Sentiment Analysis Algorithms through Azure Machine Learning: Analysis and Comparison. *Modern Applied Science*, 12(7), p.49.

Ilia Lorin. (2020). *Natural Language Processing (NLP) Use Cases in Business - MobiDev* [online]. Available from: https://mobidev.biz/blog/natural-language-processing-nlp-use-cases-business [accessed 11 April 2022].

India Population (2022) - Worldometer. (2022). *worldometers* [online]. Available from: https://www.worldometers.info/world-population/india-population/ [accessed 8 June 2022].

Inés Roldós. (2020). *Major Challenges of Natural Language Processing (NLP)* [online]. Available from: https://monkeylearn.com/blog/natural-language-processing-challenges/ [accessed 11 April 2022].

Irawaty, I., Andreswari, R. and Pramesti, D. (2020). Vectorizer Comparison for Sentiment Analysis on Social Media Youtube: A Case Study. In: *2020 3rd International Conference on Computer and Informatics Engineering, IC2IE 2020*. Institute of Electrical and Electronics Engineers Inc., pp.69–74.

Jo Hartley. (2021). *The Languages of India: What Languages are Spoken in India?* [online]. Available from: https://www.berlitz.com/blog/indian-languages-spoken-list [accessed 8 June 2022].

Kadriu, A., Abazi, L. and Abazi, H. (2019). Albanian Text Classification: Bag of Words Model and Word Analogies. *Business Systems Research*, 10(1), pp.74–87.

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(3).

Kumar, A. and Sachdeva, N. (2020). Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. In: *Multimedia Systems*. Springer.

Kumar, V. and Subba, B. (2020). A tfidfvectorizer and SVM based sentiment analysis framework for text data corpus. In: *26th National Conference on Communications, NCC 2020*. Institute of Electrical and Electronics Engineers Inc.

Matthew Johnston. (2022). *7 Companies Owned by Google's Parent Company Alphabet (GOOGL)*, [online]. Available from: https://www.investopedia.com/investing/companies-owned-by-google/ [accessed 8 June 2022].

Mundra, S. and Mittal, N. (2021). Evaluation of text representation method to detect cyber aggression in hindi english code mixed social media text. In: *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp.402–409.

Natural Language Processing - Ela Kumar - Google Books. Available from: https://books.google.ie/books?hl=en&lr=&id=FpUBFNFuKWgC&oi=fnd&pg=PP2&dq=history+of+natural+language+processing&ots=GFy26LlyPw&sig=Qw6__PkPsebesXomRymAy6PXRsI&redir_esc=y#v=onepage&q=alan&f=false [accessed 26 May 2022].

Natural Language Processing (NLP) Examples | Tableau. Available from: https://www.tableau.com/learn/articles/natural-language-processing-examples [accessed 11 April 2022].

Nguyen, T.H., Shirai, K. and Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24), pp.9603–9611.

Pamela Fox. (2018). *Expressing an algorithm | AP CSP (article) | Khan Academy* [online]. Available from: https://www.khanacademy.org/computing/ap-computer-science-principles/algorithms-101/building-algorithms/a/expressing-an-algorithm [accessed 11 April 2022].

Qu, S., Yang, Y. and Que, Q. (2021). Emotion classification for spanish with xlm-roberta and textcnn. In: *CEUR Workshop Proceedings*. CEUR-WS, pp.94–100.

Rachel Wolff. (2020). *7 Benefits of Natural Language Processing (NLP)* [online]. Available from: https://monkeylearn.com/blog/nlp-benefits/ [accessed 11 April 2022].

Ripul Agarwal. (2020). *Sentiment Analysis of YouTube Comments | Analytics Steps* [online]. Available from: https://www.analyticssteps.com/blogs/sentiment-analysis-youtube-comments [accessed 11 April 2022].

Sentiment Analysis Guide. (2020). *Monkey Learn* [online]. Available from: https://monkeylearn.com/sentiment-analysis/ [accessed 8 June 2022].

Serrano-Guerrero, J., Olivas, J.A., Romero, F.P. and Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311, pp.18–38.

Shah, S.R., Kaushik, A., Sharma, S. and Shah, J. (2020). Opinion-mining on marglish and devanagari comments of youtube cookery channels using parametric and non-parametric learning models. *Big Data and Cognitive Computing*, 4(1), pp.1–19.

Shemmy Majewski. (2020). *7 Key Benefits Of Using Natural Language Processing In Business* [online]. Available from: https://dlabs.ai/blog/7-key-benefits-of-using-natural-language-processing-in-business/ [accessed 11 April 2022].

Singh, M. and Goyal, V. (2020). Sentiment Analysis of {E}nglish-{P}unjabi Code-Mixed Social Media Content. In: *Proceedings of the 17th International Conference on Natural Language Processing (ICON): System Demonstrations*. NLP Association of India (NLPAI), pp.24–25. Available from: https://aclanthology.org/2020.icon-demos.9.

Srivastava, V. and Singh, M. (2021). Quality Evaluation of the Low-Resource Synthetically Generated Code-Mixed Hinglish Text. In: *INLG 2021 - 14th International Conference on Natural Language Generation, Proceedings*.

Stephen J. Bigelow. (2022). *What Is a SWOT Analysis? Definition and Examples - TechTarget* [online]. Available from: https://www.techtarget.com/searchcio/definition/SWOT-analysis-strengths-weaknesses-opportunities-and-threats-analysis [accessed 11 April 2022].

Swaminathan, S., Ganesan, H.K. and Pandiyarajan, R. (2020). HRS-TECHIE@Dravidian-CodeMix and HASOC-FIRE2020: Sentiment analysis and hate speech identification using machine learning, deep learning and ensemble models. In: *CEUR Workshop Proceedings*. CEUR-WS, pp.241–252.

Thelwall, M. (2018). Gender bias in machine learning for sentiment analysis. *Online Information Review*, 42(3), pp.343–354.

UCI Machine Learning Repository: Youtube cookery channels viewers comments in Hinglish Data Set. Available from: https://archive.ics.uci.edu/ml/datasets/Youtube+cookery+channels+viewers+comments+in+Hinglish [accessed 8 April 2022].

Uma Gunturi. (2020). *A Primer on Code Mixing & Code Switching! | by Uma Gunturi | Medium* [online]. Available from: https://umagunturi789.medium.com/a-primer-on-code-mixing-code-switching-9bbde2a15e57 [accessed 11 June 2022].

Valencia, F., Gómez-Espinosa, A. and Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy*, 21(6).

Vivek Srivastava. (2021). *A representative Hinglish sentence and the corresponding parallel... | Download Scientific Diagram* [online]. Available from: https://www.researchgate.net/figure/A-representative-Hinglish-sentence-and-the-corresponding-parallel-Hindi-English-sentences_fig1_352432102 [accessed 11 April 2022].

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A. and Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21), pp.7375–7385.

Ximena Bolaños. (2020). *Natural Language Processing with Machine Learning* [online]. Available from: https://www.encora.com/insights/natural-language-processing-with-machine-learning [accessed 11 April 2022].

YouTube | History, Founders, & Facts | Britannica. Available from: https://www.britannica.com/topic/YouTube [accessed 26 May 2022].

Appendix