**Abstract:**

One of the most influential virtual learning sources in the present world is YouTube which has been accessed by billions of Internet users. According to this scenario, the number of people who create YouTube videos popularly called YouTubers has increased. This project will be useful in analysing the YouTube comments given by users for the videos posted by YouTubers. It helps in knowing the thoughts of users according to the video content and helps YouTubers to post videos with better quality and content. As India has millions of users, the comments on Indian YouTube cookery channels are taken for analysis. Indians mostly use Mix-Code language in commenting i.e., Hinglish which is the combination of Hindi and English languages. Different Vectorization techniques using TF-IDF, Term Frequency, Count Vectorizer, Bert Transformers, 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.

**Introduction:**

YouTube is an online video-sharing social media platform that started on 14th February 2005 and is owned by Google in October 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 (YouTube - Wikipedia n.d.). The Videos on YouTube include short films, movies, documentaries, cooking channels, educational and technological related, etc. As everyone is becoming independent in their food preferences for diet control, cookery channels are very much helpful for learning food preparation. 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 get knowledge about the viewers’ thoughts on the videos, they must manually read the comments and prepare for the next video. 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.

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 - Wikipedia n.d.). Sentimental Analysis is opinion and emotion Analysis by extracting them 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 for getting the theme of the text. These words are converted to vectorized forms using different vectorization methods and trained to Machine Learning (ML) model. Generally, Classification models are integrated into the NLP process. This is because different texts should be classified based on the vectors which indirectly are sentiments. As labels will be provided for training the model, Supervised learning will be applied in this project.

Machine Learning (ML) which is a term introduced by Arthur Samuel in 1959 (Timeline of machine learning - Wikipedia n.d.) involves mainly two types of learning namely Supervised and Unsupervised. In Supervised Learning, the ML 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. 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 in the same Figure. The data consists of most of these types of comments. There are some challenges in analysing the Mix-Code languages as stop words in NLP should be given manually depending on our requirements. Some of the other Mix-Code languages (Code-mixing - Wikipedia n.d.) can be noted in Table 1.



Figure. 1. Hinglish Mix-Code Language

|  |  |
| --- | --- |
| **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

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, Data Exploration and Pre-Processing, Ethical Considerations, Future plan, and References. The problem statement, the structure of the report, research questions, and research motivation is discussed in the Introduction. The research for the hurdles faced, suggested 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 Future plan section. The work references are added in the References section.

**Research Questions:**

The Research question according to the classification of Hinglish comments using NLP and ML are mentioned below

1. Which Vectorizer is doing best during multiple vectorization techniques on Hinglish Mix-Code before Modelling?
2. To understand Feature Engineering methods of NLP.
3. To observe the performance of models through best evaluation results like Accuracy, Precision, recall, Confusion matrices, and Classification Report for this Hinglish classification?
4. To study different Cross-validation methods and Scaling techniques and examine their effect on the evaluation results of models.
5. Effect of Principal Component Analysis and Independent Component Analysis on the ML models.
6. To observe the potential of different algorithms through both parametric and non-parametric models and to investigate the prospective of transfer learning in mix-code.