**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 intention 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 | History, Founders, & Facts | Britannica 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 - Ela Kumar - Google Books 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 Natural Language Processing processes. 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 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. 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 Natural Language Processing 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 Natural Language Processing and Machine Learning are mentioned below

1. Which Vectorizer during multiple vectorization techniques is best pair for Machine Learning models on Hinglish Mix-Code?
2. Which parametric or non-parametric model gives best evaluation results like Accuracy, Precision, recall, Confusion matrices, and Classification Report for this Hinglish classification?
3. Is Principal Component Analysis and Independent Component Analysis on the Machine Learning models helps in getting good results for Mix-Code models?

**Literature Review:**

**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 analyse 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. 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.

Diagram

Description automatically generated

Figure. 2. Flow of Methodology

*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 by means of 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 equally 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 recipe and video in 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 comes under any of other labels like praising or showing gratitude or querying or about recipe or video.  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

*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.

*Data Visualization –*

The main purpose of this data visualization is to analyse the data visually. It provides the well organised visual representation of data to easily analyse and interpret the understandings. The distribution of labels, stop words, hashtags, word counts, character counts, numerical values present, etc in the data will be analysed using visualizations. This will be achieved by plotting the graphs in count plots using matplotlib or seaborn libraries.