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

Murthy S Routhula (D00243413), M.Sc. Data Analytics, Department of Computing Science and Mathematics

Introduction

YouTube is a popular learning platform that started on 14th February 2005. Different videos are being uploaded and viewed by users every day based on different topics and genres. For example, international students who have habituated to the home food learn to cook food themselves using YouTube videos. Many cooking channels have been started by YouTubers for a wide variety of cuisines. Users give their reviews based on the video content through comments. Different countries have regional languages and while commenting, people use the mix-code which is the combination of two or more languages.

Mix-Code	Languages
Poglish	Polish + English
Greeklish	Greek + English
Svorsk	Swedish + English

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

For example, India has nearly 121 languages, and Hindi is the most spoken language among them. Hinglish comments which are a mix-code of Hindi and English languages are taken for the analysis.

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 (Srivastava and Singh 2021)

Based on the number of comments per video, it is difficult for the YouTubers to manually read the comments as it is a highly time-consuming task. This project will be helpful in analyzing the mix-code comments using Sentimental Analysis. It is the Natural Language Processing technique that is helpful for determining whether the data is positive or negative or neutral (Sentiment Analysis Guide 2020).

Research Objectives & Questions

Research objectives include the investigation of multiple vectorization techniques and feature engineering methods. It is important to observe the potential of different algorithm models on the vectorized datasets of mix-code comments. Application of different cross-validation techniques while modeling is helpful in evaluation. The research questions of the project are,

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?

Methodology

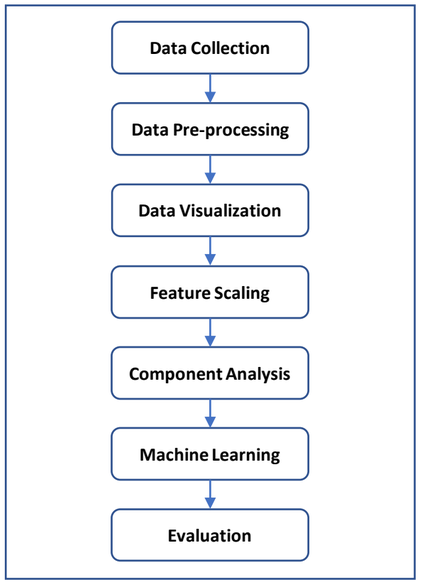

Figure.2 Flow of the Project

About Data

The data is open source and is collected from the UCI website (UCI Machine Learning Repository: Youtube cookery channels viewers comments in Hinglish Data Set n.d.).

Data Labels	Kabita's Kitchen	Nisha Madhulika
Gratitude	700	700
About Recipe	700	700
About Video	700	700
Praising	700	700
Hybrid	700	700
Undefined	700	700
Suggestion or Query	700	700

Table. 2. Labels and Distribution of Data

Data Pre-processing and Visualization

The raw data consists of many line breaks and smiley symbols. They will be removed in the preprocessing stage.

Visualization Analysis is carried out to analyze labels, stop words, hashtags, word counts, character counts, numerical values, etc. present in the data.

Vectorization techniques

- Term Frequency (TF)
- Term Frequency – Inverse Document Frequency (TF-IDF)
- Count Vectorizer (CV)
- Bidirectional Encoder Representations from Transformers (BERT)
- Generative Pre-trained Transformer (GPT)
- Cross-Lingual Language Model (XLM)

Feature Scaling

- Min-Max Scaling
- Standard Scaling
- Normalize Scaling
- Binary Scaling

Principal Component Analysis (PCA) and Independent Component Analysis (ICA)

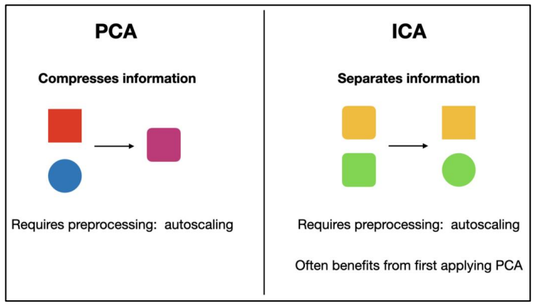

Figure. 3. PCA and ICA Explanation (Shawhin 2021)

Modeling

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. 3. Machine Learning Models

Cross-validation methods

- Train-test split
- Random train-test split
- K-fold
- Leave one out

Evaluation & Results

As this study is based on Supervised learning of classification type, the following evaluation metrics are to be compared between both parametric and non-parametric models.

- Accuracy
- Precision
- Recall
- F1-Score
- Classification Report
- Confusion matrix
- Area Under Curve (AUC)

Evaluation metrics are based on the True positives, True negatives, False positives, and False negatives after testing the model. The results obtained after modeling are compared based on different combinations of scaling, component analysis, and cross-validations.

Ethical Considerations

- Data collection and usage.
- Data storage, security, and stewardship.
- Data hygiene and relevance.
- Identifying and addressing harmful bias
- Validation and testing of models
- SWOT Analysis

Conclusion and Future work

As YouTube is one of the popular mediums for free learning, many people prefer to learn and try new cuisines and recipes from it. So, the quality of content that is uploaded by YouTubers is important according to the users' reviews. This use case is very helpful for the cooking channel admins in creating content based on users' requirements. Also, the main aim of this sentimental analysis is to find the best model for the Hinglish mix-code comments based on evaluation metrics.

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 with different mix-codes will be covered for the sentimental analysis.

Video Category	Average views a single video uploaded in this category gets in India
Entertainment	10K views
How to and Style	7.9K views
Pets and Animals	7.2K views
Science and technology	6.9K views
Automobiles	6.1K views
Education	5.1K views
Gaming	2.9K views
Travel	2.9K views
Vlogs	2.6K views

Figure. 4. YouTube Video Categories and Views (Adgully 2018)

References

- Independent Component Analysis (ICA) | by Shawhin Talebi | Towards Data Science. (n.d.). Retrieved June 14, 2022, from https://towardsdatascience.com/independent-component-analysis-ica-a3eba0ccec35
- Native content on YouTube a big hit with Indian viewers: Vidooly report. (n.d.). Retrieved June 14, 2022, from https://www.adgully.com/native-content-on-youtube-a-big-hit-with-indian-viewers-vidooly-report-78866.html
- Srivastava, V., & Singh, M. (2021). Quality Evaluation of the Low-Resource Synthetically Generated Code-Mixed Hinglish Text. INLG 2021 - 14th International Conference on Natural Language Generation, Proceedings.
- Uma Gunturi. (2020). A Primer on Code Mixing & Code Switching! | by Uma Gunturi | Medium. https://umagunturi789.medium.com/a-primer-on-code-mixing-code-switching-9bbde2a15e57