Analysing The Sentiments Of Marathi-English Code-Mixed Social Media Data Using Machine Learning Techniques

Varad Patwardhan
Computational Linguistics Research Lab
Pune Institute of Computer Technology
Pune, India
varadp2000@gmail.com*

Gauri Takawane
Computational Linguistics Research Lab
Pune Institute of Computer Technology
Pune, India
gauri.takawane@gmail.com

Nirmayi Kelkar
Computational Linguistics Research Lab
Pune Institute of Computer Technology
Pune, India
nimu.kelkar@gmail.com

Omkar Gaikwad
Computational Linguistics Research Lab
Pune Institute of Computer Technology
Pune, India
omkargaikwad9552@gmail.com

Rutwik Saraf
Computational Linguistics Research Lab
Pune Institute of Computer Technology
Pune, India
rutwiksaraf3@gmail.com

Sheetal Sonawane
Computational Linguistics Research Lab
Pune Institute of Computer Technology
Pune, India
sssonawane@pict.edu

***Abstract*—A vast amount of data is generated every day through social media platforms. Various techniques and methodologies are used to bring different forms of data to use. One such form of data is textual data generated from social media platforms in the form of chats, comments, and tweets. The term "code-mixed data" describes data that combines components of different languages or linguistic subgroups such as text written in several different languages or speech that shifts between languages. Due to increased social media use and worldwide communication, many individuals are using multiple languages in their daily communication, making this type of data even more crucial. Machine translation, speech recognition, and text categorization are just a few examples of natural language processing activities that can be performed on code-mixed data. Research on code-mixed data can also aid in the understanding of multilingual communication. In this paper, we present an empirical study on the problem of word-level language identification and text normalisation for Marathi-English code-mixed text. We have created a new dataset of 1009 sentences that exhibit code-mixing of Marathi (Romanised) and English textual data. This data was collected from Whatsapp chats and Youtube comments.**

***Index Terms*—NLP, SVM, Naive Bayes, Code-mixed Marathi English Romanised, sentiment analysis, language annotations**

I. Introduction

Marathi is an Indo-Aryan language that is spoken by the Marathi people. It has 90 million speakers globally. It has the third largest number of speakers in India, after Hindi and Bengali according to the Census Table of India[1]. It is also the official language of the states of Maharashtra and Goa. Many people use Marathi on social media platforms like WhatsApp and Facebook. As most of the population presently is bilingual or multilingual, people tend to write in their preferred language. Code-mixing is a linguistic phenomenon where multiple languages are used in the same sentence. This form of communication is common in multilingual societies such as when Marathi is code-mixed with English. Due to ease of the font, people prefer writing in the Roman script rather than the Devanagari script. To understand and analyse the sentiments of people on social media, it is necessary to understand code-mixed data. This paper aims to develop a deeper understanding of Marathi-English code-mix data and provide resources to any fellow researcher wanting to explore this domain further.

[1]Census Table of India- https://censusindia.gov.in/census.website/data/census-tables

II. Literature Survey

Sentiment analysis is an important application in NLP and Information Retrieval. Recent studies on sentiment analysis have shown remarkable achievements in making useful decisions [8]. NLP algorithms have proved to be valuable in the COVID-19 pandemic to understand the sentiments of the people [12].

Various studies have been conducted on sentiment analysis of code-mixed data. Studies have been performed on Chinese and English code-mixed data [11] where translation and non-translation-based techniques were used for text analysis. Work is being done for the normalization of Konkani-English data pairs [18] where a shallow parser is created for data mining and sentiment analysis [17]. Various parameters are taken into account for the analysis of sentiment [16]. Preprocessing techniques have been used for the structuring of data such as POS tagging [20]. Competitions have been conducted where participants received annotated datasets and created models for classification [19] and sentiment analysis of tweets

[15]. Another application of sentiment analysis is Offensive Language Detection [1] using BERT, LSTM, and GRU [7].

Platforms such as Twitter provide huge amounts of text data in the form of tweets. This data has been used to create Probabilistic Latent Semantic Analysis via SVM [6]. Data should be in proper labelled form for models to work efficiently, hence various forms of the corpus are created such as a Tamil-English corpus [3] and an annotated Telugu-English corpus [14]. Apart from machine learning techniques, ensemble learning techniques are used for sentiment prediction [10]. Another model named SentiWordNet is used but has a decline in performance compared to SVM models [5].

A literature survey indicated a model named VADER that has a better performance than SentiNetWord [21]. According to our study, SVM is used more often as compared to other models [6] such as LSTM [9] and Naïve Bayes [2]. A triple-loss function employed in the BERT model improves its efficiency [7]. As per our knowledge, no work has been conducted on code-mixed data of Marathi-English languages which creates a gap in the analysis of texts that are Romanized Marathi data pairs.

III. Dataset Creation

Dataset is manually curated from social media platforms like WhatsApp and Youtube. The scraped data is generated from WhatsApp chats and YouTube comments in the time-frame from year 2021–2022. The dataset covers sentences from the education, entertainment and academics domain. The dataset is publicly available on github.[2]

The sentences in the dataset were preprocessed by removing special characters like *, @, #, etc, emojis, and punctuations. Additionally, they were manually annotated with language identification tags, POS (Part of Speech) tagging, and sentiment analysis.

A. Language Identification

We have considered the following tags for Language Identification:

- EN - for a completely English word. e.g: friend
- MR - for a completely Marathi word. e.g: mitra
- BO - for words that can be considered in both English and Marathi. e.g: Simran
- UN - for words that do not belong to any of the above categories. e.g: aankhein

B. Parts of Speech Tagging

We have classified each token as follows for Parts of Speech tagging:

- NOU - for Noun
- VB - for Verb
- ADV - for Adverb
- ADJ - for Adjective
- UN - for other parts of speech like pronouns, conjunctions, etc.

[2]Code Mixed Marathi English Dataset

C. Sentiment Tagging

We have categorized each sentence as follows for Sentiment Tagging:

- Positive
- Negative
- Neutral

D. Dataset Statistics

Dataset statistics are shown in the tables below:

TABLE I
Language Identification

Tokens	Quantity
MR	8203
EN	3308
BO	482
UN	123
$TOTAL$	12116

Table I shows token-based language identifications together with the corresponding frequency count. Tokens in Marathi are denoted by MR, those in English by EN, those used in both languages by BO, and all other tokens by UN.

TABLE II
Part of Speech

Tokens	Quantity
ADV	694
VB	2858
NOU	2686
ADJ	1052
UN	4826
$TOTAL$	12116

Table II demonstrates the identification of parts of speech using tokens and their frequency count. Adjectives are referred to as ADJ, Adverbs as ADV, Verbs as VB, Nouns as NOU, and Adverbs as ADJ. All other tokens are represented by UN.

TABLE III
Sentiments

Sentiment	Quantity	Percentage
$Positive$	392	38.85%
$Negative$	297	29.44%
$Neutral$	320	31.71%

The number of positive, neutral, and negative texts in the dataset is shown in Table III.

IV. Methodology

A. Part of speech tagging

A fundamental level of syntactic analysis for a specific word or sentence is provided by the part of speech. The task of

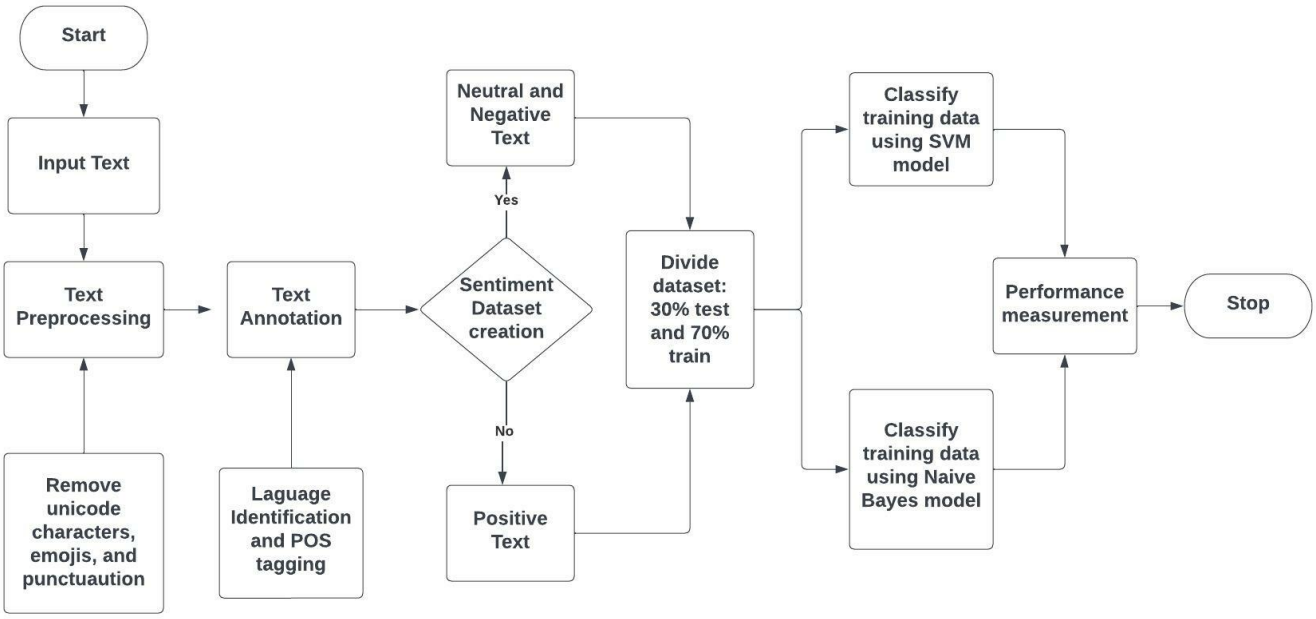

Fig. 1. System Flow diagram for code-mix analysis

labelling each word in a sentence with the correct part of speech is known as "POS tagging". The emphasis of this paper has been on nouns, verbs, adverbs, and adjectives. Parts of the speech were tagged manually for both English and Marathi words. The following guidelines were used while tagging the words:

1) Words were classified according to their context because each context affects how a word functions.
2) Regardless of the POS tag assigned to the word in its native language, words that were embedded in sentences in another language were labelled according to the context of the prevailing language.
3) Adverbs (ADV), Adjectives (ADJ), Verbs (VB), Nouns (NOU), and the remaining words as Unknown (UN)

B. Language Identification

In code-mixed languages, language identification is crucial for separating words from one language from another. Language identification consists of simply labelling words with the appropriate language tag, which aids in differentiating between languages even when the words have the same spelling. For both English and Marathi words, language identification was done manually. While tagging the terms, the following rules were used:

1) The symbol 'MR' was used to identify Romanized Marathi words.
2) The symbol 'EN' was used to mark English words.
3) The symbol 'BO' (Both) was used to identify words that are used often in both languages.
4) The Symbol 'UN' (Unknown) was used to designate words that didn't fit into any of these categories.

C. Sentiment Analysis

Identification, extraction, and analysis of subjective data are all part of sentiment analysis. It uses computational linguistics, text analysis, natural language processing, etc. Sentiment analysis is commonly used to establish people's opinions on a variety of subjects, such as reviews, survey findings, online and social media posts, etc. Each statement's emotions were given their own tag. The sentences were tagged in accordance with the following rules:

1) If the response was extremely positive, the number 1 was assigned.
2) In the event of a strong negative response, -1 was given.
3) 0 was assigned in the event of a neutral response.
4) In the event of a mixed response, 0 or the dominating emotion was assigned.

D. Algorithms

1) SVM: A supervised machine learning approach called the Support Vector Machine (SVM) is employed for classification or regression. A hyper-plane that distinguishes between the various data categories is found via SVM. The text may be classified as having a good, neutral, or negative attitude using the hyperplane.[4]

2) Naïve Bayes: A series of classification techniques based on the Bayes theorem are collectively referred to as "Naive Bayes classifiers." It is not one algorithm, but rather a collection of algorithms that share the core tenet that every pair of characteristics that is categorised stands alone from the rest. By taking into account the chance of an earlier event occurring, the Bayes theorem determines the likelihood of an event occurring.[13]

E. Methodology

Romanized Marathi does not currently have a pre-trained model. As a result, for the dataset evaluation, we chose to employ SVM and Naive Bayes, two machine learning algorithms. Emojis and punctuation were removed from the data as part of the pre-processing procedure and sentiment analysis was then carried out with the help of SVM and Naive Bayes. During this analysis, we only looked at sentences and ignored word language or word parts of speech. Accuracy may be further improved by utilizing feature extraction and taking these factors into account.

Emojis, punctuation, links, new line characters, and spaces were first removed from the collected dataset using regex at the beginning of the processing procedure. The Emoji package was utilised for emoji removal. After that, the dataset was divided into a 70:30 train and test split. The train segment had 706 sentences, and the test part had 303 sentences. The dataset was subsequently vectorized in TF-IDF format using feature extraction. After that, SVM and Naive Bayes models were trained on the training dataset. The test portion was utilised to construct the confusion matrix and to assess the F1 score as well as the accuracy.

Naive Bayes:All the sentences that made up the corpus were made to go through the preprocessing steps of tokenization, stemming, POS tagging and lemmatization to form a vector of terms. Each of the sentences in the training set were manually labelled as positive, negative or neutral. Given a new sentence from the test data, it was also made to go through the pre-processing steps and converted to a term vector of all the terms in the corpus. Mathematically, we are interested in labelling this vector of terms t1,t(2,)...tn into a class Ci that can be either positive, negative or neutral. Thus, to find the probability - P(Ci/t1 t2...tn) Using the Naïve Bayes assumption, we can assign a class C to the sentence of terms t1,t(2,)...tn by the formula

$$C = \underset{i}{\operatorname{argmax}} \left[P(C_i) \times \prod_{j=1}^{no\ of\ terms\ in\ corpus} P(t_j/C_i) \right]$$

$$C_i \in \{\text{Positive,Negative,Neutral}\}$$

Fig. 2. Naive Bayes Formula

SVM: All the sentences that made up the corpus were made to go through the preprocessing steps of tokenization, stemming, POS tagging and lemmatization to form a vector of terms. An input document term matrix, X was prepared. SVM is primarily a binary classification technique in which a maximum margin classifier is built up that "learns" W and b matrices such that the function.

V. Results

The test data was used to calculate the metrics of the SVM and Naive Bayes models after training them on the training set of data. Results were calculated for the confusion matrix, accuracy score, and f1 scores. A summary of the findings can be found in tables IV and V.

$$L = \frac{1}{2} W^T W + C \sum_{i=1}^{N} max[0,1 - \xi_i]$$

$$\xi_i = y_i(X_i W + b)$$

Fig. 3. SVM Formula

TABLE IV
SVM Confusion Matrix

Metrics	Positive	Neutral	Negative
Positive	40	33	16
Neutral	20	39	37
Negative	15	26	77

TABLE V
Naive Confusion Matrix

Metrics	Positive	Neutral	Negative
Positive	33	31	25
Neutral	20	36	40
Negative	20	23	75

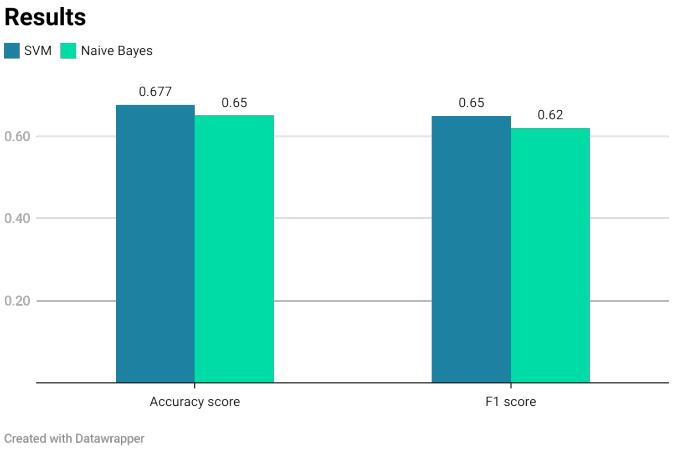

Fig. 4. Results

The confusion matrix for the SVM-based model is shown in Table IV, and the confusion matrix for the Naive Bayes model is shown in Table V. Figure 4 illustrates how SVM and Naive Bayes compare in terms of F1 score and Accuracy. The SVM appears to perform better than the Naive Bayes model in terms of results. In comparison to neutral sentiment, the accuracy and F1 score of positive and negative statements are higher. After including more neutral sentences in the dataset and extracting more features thoroughly, these results may

be improved. SVM surpasses Naive Bayes in this instance, however their approaches are very different. Therefore, by choosing various features, we may be able to improve the result of the Naive Bayes model.

VI. Conclusion and Future Scope

The analysis of sentiment of code mixed data of Marathi and English language is explored in this research work. As per the authors' knowledge this is the first attempt to handle sentiments of code mixed Marathi and English languages. Using more thorough feature extraction and expanding the dataset, the accuracy and F1 score may be improved. In addition, deep learning Models such as BERT and RoBERTa can be used for better results. A generalized model of processing and analyzing the Indian language can be explored further using the Graph model.

Acknowledgment

This work was carried out in the Computational Linguistics Research Laboratory, PICT. We are thankful to the institute for providing all the required help and support to carry out this research work.

References

[1] Kalaivani Adaikkan and Thenmozhi Durairaj. "Multilingual Hate Speech and Offensive Language Detection in English Hindi and Marathi languages". In: *Fire*. 2021.

[2] Mohammed Arshad Ansari and Sharvari Govilkar. "Sentiment analysis of mixed code for the transliterated hindi and marathi texts". In: *International Journal on Natural Language Computing (IJNLC) Vol* 7 (2018).

[3] Bharathi Raja Chakravarthi et al. "Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text". In: (2020). DOI: 10.48550/ARXIV.2006.00206. URL: https://arxiv.org/abs/2006.00206.

[4] Vladimir Cortes Corinna and Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[5] Mohammad Fikri and Riyanarto Sarno. "A comparative study of sentiment analysis using SVM and Senti Word Net". In: *Indonesian Journal of Electrical Engineering and Computer Science* 13 (Mar. 2019), pp. 902–909. DOI: 10.11591/ijeecs.v13.i3.pp902-909.

[6] Han et al. "Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter DataSet". In: *Applied Sciences* 10 (Feb. 2020), p. 1125. DOI: 10.3390/app10031125.

[7] Adeep Hande et al. "Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages". In: *arXiv preprint arXiv:2108.03867* (2021).

[8] Vishal A. Kharde and S.S. Sonawane. "Article: Sentiment Analysis of Twitter Data: A Survey of Techniques". In: *International Journal of Computer Applications* 139.11 (Apr. 2016). Published by Foundation of Computer Science (FCS), NY, USA, pp. 5–15.

[9] Arouna Konate and Ruiying Du. "Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques". In: *Wuhan University Journal of Natural Sciences* 23.3 (2018), pp. 237–243.

[10] SR Mithun Kumar et al. *An ensemble model for sentiment classification on code-mixed data in Dravidian Languages*. Tech. rep. EasyChair, 2021.

[11] Kong Lim and Tong Lim. "A Review on Sentiment Analysis for Code-Mix Chinese and English Text on Social Media". In: Jan. 2020, pp. 53–57. DOI: 10.56453/icdxa.2020.1001.

[12] Aboli Marathe et al. "Leveraging Natural Language Processing Algorithms to Understand the Impact of the COVID-19 Pandemic and Related Policies on Public Sentiment in India". In: *2021 International Conference on Communication information and Computing Technology (ICCICT)*. 2021, pp. 1–5. DOI: 10.1109/ICCICT50803.2021.9510070.

[13] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. "Spam Filtering with Naive Bayes - Which Naive Bayes?" In: Jan. 2006.

[14] Pruthwik Mishra, Prathyusha Danda, and Pranav Dhakras. "Code-mixed sentiment analysis using machine learning and neural network approaches". In: *arXiv preprint arXiv:1808.03299* (2018).

[15] Braja Gopal Patra, Dipankar Das, and Amitava Das. "Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017". In: *arXiv preprint arXiv:1803.06745* (2018).

[16] Haiyun Peng, E. Cambria, and Amir Hussain. "A Review of Sentiment Analysis Research in Chinese Language". In: *Cognitive Computation* 9 (2017), pp. 423–435.

[17] Akshata Phadte and Radhiya Arsekar. "Part-of-Speech Tagger for Konkani-English Code-Mixed Social Media Text". In: Jan. 2018, pp. 303–307. ISBN: 978-3-319-91946-1. DOI: 10.1007/978-3-319-91947-8_31.

[18] Akshata Phadte and Gaurish Thakkar. "Towards normalising Konkani-English code-mixed social media text". In: *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*. 2017, pp. 85–94.

[19] Deepesh Sharma. "TADS@Dravidian-CodeMix-FIRE2020: Sentiment Analysis on CodeMix Dravidian Language". In: *Fire*. 2020.

[20] Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. "A Twitter Corpus for Hindi-English Code Mixed POS Tagging". In: Jan. 2018, pp. 12–17. DOI: 10.18653/v1/W18-3503.

[21] C Tho et al. "Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach". In: *Journal of Physics: Conference Series*. Vol. 1869. 1. IOP Publishing. 2021, p. 012084.