Sentiment Analysis of Marma Language Using Machine Learning Techniques

Course Code: CSE478
Course Title: Project/Thesis

Submitted By:
Name: Mohammad Nimour Hossain
ID: 20CSE010
Name: Iftaker Siddique
ID: 20CSE036
Session: 2020-21

Supervised By:
Dr. Mrinal Kanti Baowaly
Associate Professor,
Department of Computer Science and Engineering,
Gopalganj Science and Technology University.

Department of Computer Science and Engineering Gopalganj Science and Technology University

Statement of Originality

We hereby declare that this thesis is our own work and that neither this thesis nor any part of it has been submitted elsewhere for the award of any degree or diploma.

…… ……

Mohammad Nimour Hossain
Date:
ID: 20CSE010
4th Year 1st Semester.

Iftaker Siddique
Date:
ID: 20CSE036
4th Year 1st Semester.

Thesis Approval

This Thesis is submitted to the Department of Computer Science and Engineering, Gopalganj Science and Technology University, for the partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

Supervisor Approval

……

Dr. Mrinal Kanti Baowaly
Associate Professor, Department of Computer Science and Engineering

Gopalganj Science and Technology University, Gopalganj, Bangladesh

Acknowledgment

We would like to express our deepest sense of gratitude to our respected teacher and supervisor, **Dr. Mrinal Kanti Baowaly**, Associate Professor, Department of Computer Science and Engineering (CSE), Gopalganj Science and Technology University (GSTU). We are grateful for his untiring guidance, constant supervision, enthusiastic encouragement, and sagacious advice throughout the entire period of our project.

We sincerely thank him for his patience and for giving his time so generously to guide us, which has been very much appreciated.

Mohammad Nimour Hossain
20CSE010

Iftaker Siddique
20CSE036

Abstract

The Marma language, spoken by approximately 200,000 people in the Chittagong Hill Tracts, faces a significant digital divide due to the scarcity of computational resources. Although native speakers often use Bengali characters for digital communication, existing Natural Language Processing (NLP) tools designed for Bengali are ineffective for Marma due to distinct grammatical structures. This thesis presents a machine learning-based approach for Sentiment Analysis of the Marma Language to classify text into Positive, Negative, and Neutral categories.

To address the lack of data, a primary dataset is being developed by collecting sentence pairs directly from native speakers. The methodology involves rigorous preprocessing, including tokenization, stopword removal, and N-gram vectorization, followed by a comparative analysis of four supervised learning models: Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree. This study aims to identify the most accurate classifier, establishing a critical baseline for future research and supporting the digital inclusion of the Marma community.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Background

The Marma language is spoken by approximately 200,000 people in the Chittagong Hill Tracts. It has its own script, but in the digital era, it is an under-resourced language. Most native speakers use Bengali characters to communicate on social media, creating a unique linguistic challenge. This thesis aims to introduce machine learning techniques to preserve and process the Marma language.

1.2 Problem Statement

There is currently no standard dataset or digital tool for Marma sentiment analysis. Existing NLP tools designed for the Bengali language do not work for Marma due to distinct grammatical structures and vocabulary. This creates a "digital divide" where the Marma community cannot access modern AI benefits.

1.3 Objectives

The primary objectives of this research are:

- To create a primary dataset of Marma sentences (labeled Positive, Negative, Neutral).
- To apply preprocessing techniques: Data Cleaning, Tokenization, Stopword Removal, and N-gram Vectorization.
- To implement and compare four Machine Learning models: Logistic Regression, SVM, KNN, and Decision Tree.
- To identify the most accurate model for Marma text classification.

1.4 Scope of the Study

The study is limited to the sentiment classification of Marma text written in Bengali characters. It focuses on supervised machine learning techniques rather than deep learning due to the current dataset size.

1.5 Significance of the Study

This work establishes the first baseline for Marma NLP, providing a dataset and a comparative study of models that will guide future research and support the digital inclusion of indigenous communities.

Chapter 2: Literature Review

2.1 Machine Learning in NLP

Sentiment analysis relies heavily on supervised learning. While Deep Learning is popular, traditional algorithms like SVM and Logistic Regression are often more effective for low-resource languages with smaller datasets.

2.2 Existing Studies

We reviewed sentiment analysis works in related regional languages:

- Bengali: Md. Saiful Islam (2016) – Provided frameworks for preprocessing South Asian languages.
- Hindi: Sheetal Sharma (2018) – Demonstrated ML effectiveness on Hindi text.
- Urdu: Mohammad Abid Khan (2017) – Analyzed sentiment in Urdu.
- Burmese: Yu Mon Aye (2018) – Relevant as Burmese shares linguistic roots with Marma

2.3 Research Gap

Despite these studies, resources and existing literature for the Marma language remain significantly limited. This thesis addresses this deficiency by developing a foundational dataset and analysis framework.

Chapter 3: Methodology

3.1 Introduction

Our proposed methodology consists of Data Collection, Preprocessing, Model Design and Training and Comparative Analysis.

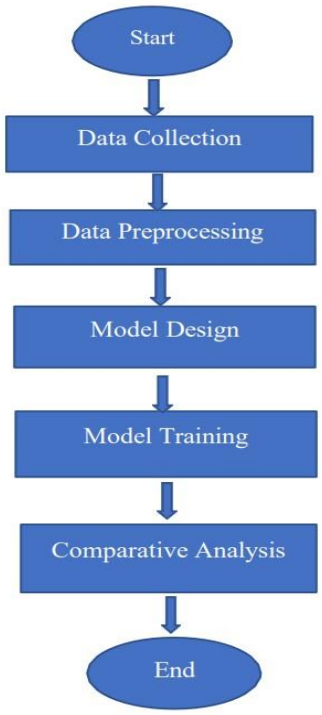

Figure 1: Methodology Diagram

3.2 Dataset Description

Since no public data exists, we are creating a Primary Dataset. We are physically collecting sentence pairs from native Marma students and speakers. Currently, 200+ sentence pairs have been collected and labeled (Positive/Negative/Neutral).

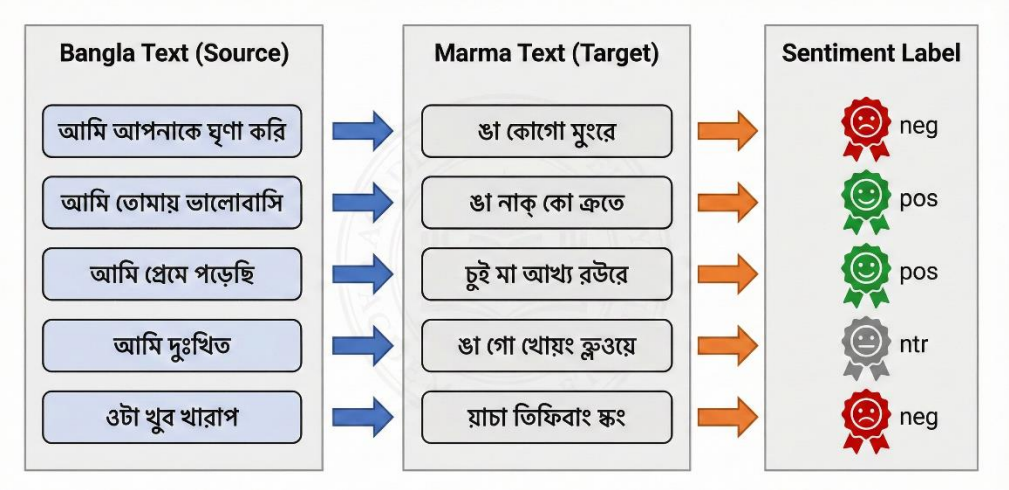

Figure 2: Sample Data Instances with Sentiment Labels

3.3 Data Processing

- Cleaning: Removing punctuation and special characters.
- Tokenization: Splitting sentences into words.
- Stopword Removal: Filtering out common non-emotional words.
- Vectorization: Using N-gram methods to convert text into numerical features.

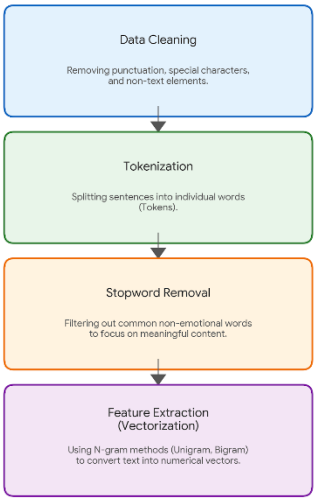

Figure 3 : The Data Preprocessing Pipeline proposed for Marma Sentiment Analysis.

3.3.1 Data Cleaning

This is the first preprocessing step aimed at reducing noise in the raw dataset. It involves systematically removing non-essential characters such as punctuation marks (e.g., `?`, `!`, `,`), special symbols (e.g., `@`, `#`), and emojis. This standardization ensures that the model focuses only on the textual content.

Figure 4 : Illustration of the Data Cleaning Process

3.3.2 Tokenization

Tokenization is the process of breaking down complex sentences into their smallest constituent units, known as "tokens." In this context, it involves splitting sentences into individual words. For instance, the sentence "Text processing is fun" would be split into the list `["Text", "processing", "is", "fun"]`.

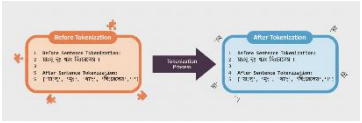

Figure 5 : Illustration of Sentence Tokenization Process

3.3.3 Stopword Removal

This step involves filtering out high-frequency words that carry little unique semantic meaning, such as "is," "the," "and," or prepositions. By removing these "stopwords," the dataset size is reduced, and the model can focus on the more meaningful, content-rich words (often the emotional or descriptive adjectives/verbs).

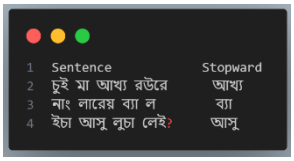

Figure 6 : Illustration of Stopwords Removal Process

3.3.4 Vectorization (N-gram)

Machine learning models cannot understand raw text, so this step converts the cleaned tokens into numerical values. Using **N-gram methods** (like Unigram or Bigram), the text is transformed into numerical feature vectors. This captures not just individual words, but potentially the context of adjacent words, turning them into mathematical input for the algorithm.

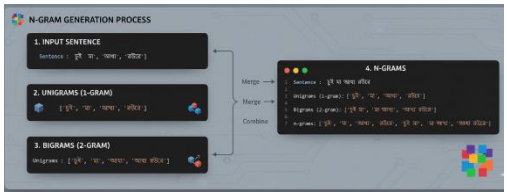

Figure 7 : Illustration of Word-to-Vector (N-gram) Process

3.4 Model Architectures

We are implementing four specific models:

- Logistic Regression: A baseline classifier.
- Support Vector Machine (SVM): Effective for high-dimensional text data.
- K-Nearest Neighbors (KNN): Classifies based on feature similarity.
- Decision Tree: A rule-based classifier.

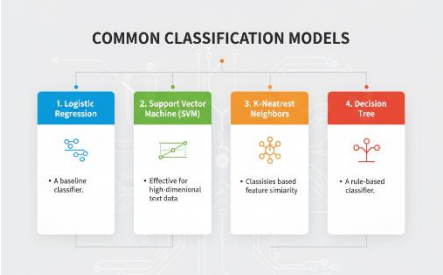

Figure 8 : Overview of Selected Machine Learning Classifiers

3.5 Training Configuration

The dataset will be split into Training (80%) and Testing (20%) sets to validate the models.

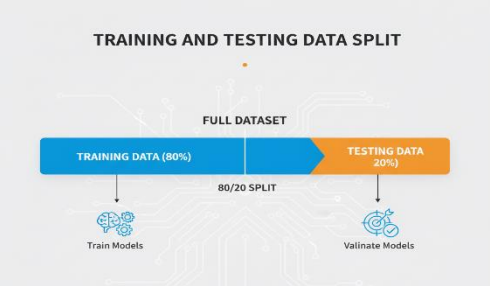

Figure 9 : Illustration of the 80/20 Train-Test Data Split.

Chapter 4: Preliminary Implementation and Expected Results

4.1 Current Implementation Status

As the thesis is ongoing, we have successfully designed the preprocessing pipeline and model architecture. To validate our code and flow, we have performed initial testing using a Bengali dataset as a reference point, which confirmed that our N-gram vectorization and model calls are functioning correctly.

4.2 Expected Accuracy and Loss

Once the full Marma dataset is collected and trained, we expect to generate the following:

- **Logistic Regression:** Expected to provide a solid baseline accuracy.
- **SVM:** Anticipated to yield the highest accuracy due to its robustness with text data.
- **KNN & Decision Tree:** Will be evaluated for comparison, though likely to be slightly less accurate than SVM on sparse text.

4.3 Comparative Analysis

We will conduct a comparative analysis using Accuracy and F1-Score to identify the most effective classifier. The F1-Score will be prioritized to balance precision and recall, ensuring the selection of the optimal model for the dataset.

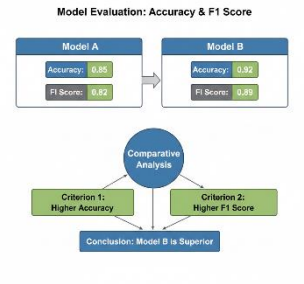

Figure 10 : Comparative Analysis of Model Performance using Accuracy and F1 Scores

Chapter 5: Discussion

5.1 Overview

The primary challenge identified during this ongoing work is the manual collection of data, which is time-consuming but necessary.

5.2 Comparative Model Suitability

Based on our literature review and preliminary tests, SVM is currently hypothesized to be the most suitable model for this task because it handles the "curse of dimensionality" in text vectors better than KNN.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

This report outlines the ongoing development of a Sentiment Analysis system for the Marma language. We have successfully defined the research gap, selected the appropriate machine learning algorithms, and initiated the data collection process. The successful implementation of this project will result in the first digital tool capable of understanding Marma sentiment.

6.2 Completed Work Summary

- Identified the research gap and reviewed relevant literature on regional sentiment analysis to establish a framework.
- Designed the complete pipeline and selected four algorithms: Logistic Regression, SVM, KNN, and Decision Tree.
- Implemented and verified the code for data cleaning, tokenization, and N-gram vectorization.
- Validated the machine learning workflow using a reference Bangla dataset to ensure technical correctness.
- Collected and labeled an initial batch of 200+ Marma sentence pairs from native students.

6.3 Pending Work

- Expand the primary dataset to thousands of sentences through continued collection from native speakers.
- Manually annotate all new data with accurate sentiment labels (Positive, Negative, Neutral).
- Train the prepared models on the full Marma dataset to generate final accuracy results.
- Perform a comparative analysis to identify the best-performing classifier based on standard metrics.Generate confusion matrices and finalize the thesis documentation and report.

References

Bengali Sentiment Analysis (Islam & Das, 2016)

Pang, B., & Lee, L. (2002). Sentiment classification using machine learning techniques. EMNLP. Das, A., & Bandyopadhyay, S. (2010). SentiWordNet for Indian Languages. AFNLP. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool. Start, M., & Kotsiantis, S. (2007). Review of classification techniques. Informatics, 31, 249–268.

Hindi Sentiment Analysis (Sharma & Jain, 2018)

Joshi, A., Balamurali, A. R., & Bhattacharyya, P. (2010). Fallback strategy for Hindi sentiment analysis. ICON. Arora, P., & Arora, P. (2015). Mining sentiment from Hindi text. International Journal of Computer Applications, 120(9), 23–26. Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0. LREC'10. Bakliwal, A., Arora, P., & Varma, V. (2012). Hindi subjective lexicon. LREC'12.

Urdu Sentiment Analysis (Khan & Amjad, 2017)

Mukund, S., & Srihari, R. K. (2010). SVM approach to Urdu sentiment analysis. AND'10. Rehman, Z., & Bajwa, I. S. (2016). Lexicon-based Urdu sentiment analysis. INTECH. Joachims, T. (1998). Text categorization with support vector machines. ECML. Syed, A. Z., Aslam, M., & Martinez-Enriquez, A. M. (2010). Lexicon-based Urdu sentiment analysis. ITAB.

Arabic Sentiment Analysis (Al-Sallab et al., 2015)

Badaro, G., Baly, R., Hajj, H., Habash, N., & El-Hajj, W. (2014). Arabic sentiment lexicon for opinion mining. EMNLP Workshop. Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Recursive autoencoders for predicting sentiment. EMNLP. Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. (2011). OCA: Opinion corpus for Arabic. JASIST, 62(10), 2045–2054. Abdul-Mageed, M., Diab, M., & Kübler, S. (2014). SAMAR: Subjectivity and sentiment analysis for Arabic social media. Computer Speech & Language, 28(1), 20–37.