International Institute of Information Technology
Bangalore

AI 829

NATURAL LANGUAGE PROCESSING

Mandate 1 (Updated)

Hate and Offensive Speech Detection in Hindi and Marathi Social Media Texts

MT2023158 Manasi Purkar
MT2023051 Sanket Patil

Index

Contents

INTRODUCTION

The widespread use of social media in today's digital age has given rise to the issue of hate speech, characterized by the use of offensive language with the intent to harm or provoke individuals or groups based on attributes such as race, religion, ethnicity, disability, or gender. With the surge in online activities, detecting and filtering out harmful content has become crucial. This project focuses on addressing this challenge in the context of two major Indian languages, Hindi and Marathi, which are widely spoken and used in social media. The goal is to develop effective hate speech detection models tailored to the linguistic and cultural nuances of these languages, given the increasing prevalence of online activities in these linguistic communities.

PROBLEM STATEMENT

As the use of social media platforms continues to grow, so does the incidence of harmful content, including hate speech, posted online. This project aims to tackle the issue of hate speech in the Hindi and Marathi languages, two prominent languages spoken in India. The lack of attention to low-resource languages like Hindi and Marathi in the domain of hate speech detection highlights a gap in research. The project leverages deep learning approaches, to classify text as hate or non-hate. The datasets used are sourced from the HASOC shared task, focusing on Twitter posts in Hindi and Marathi. The dataset consists of binary labels. We are going to implement cross lingual transfer learning approach to improve performance of target languages with limited training data. The ultimate objective is to provide effective tools for automatically identifying and mitigating hate speech in online content written in Hindi and Marathi

USE CASES

Social Media Moderation: Implementing the developed hate speech detection models on social media platforms can automate content filtering in Hindi and Marathi, improving user experience and fostering a safer online Environment.

Community Safety: Hate speech detection can empower local forums and online spaces in Hindi and Marathi to proactively address offensive language, promoting respectful discourse and ensuring the safety of community Members.

Legal Compliance: Governments and regulators can use hate speech detection technology to monitor online platforms for compliance with hate speech laws in Hindi and Marathi, enabling timely interventions and enforcement actions.

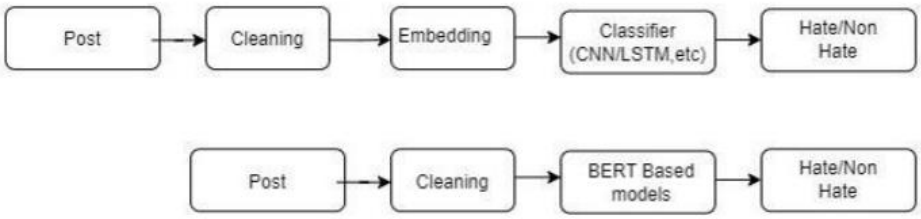
TECHNOLOGIES USED

- Python
- NLTk, Transformers
- Scikit-learn, TensorFlow
- FastText
- Pre-trained modes (e.g. mBERT)

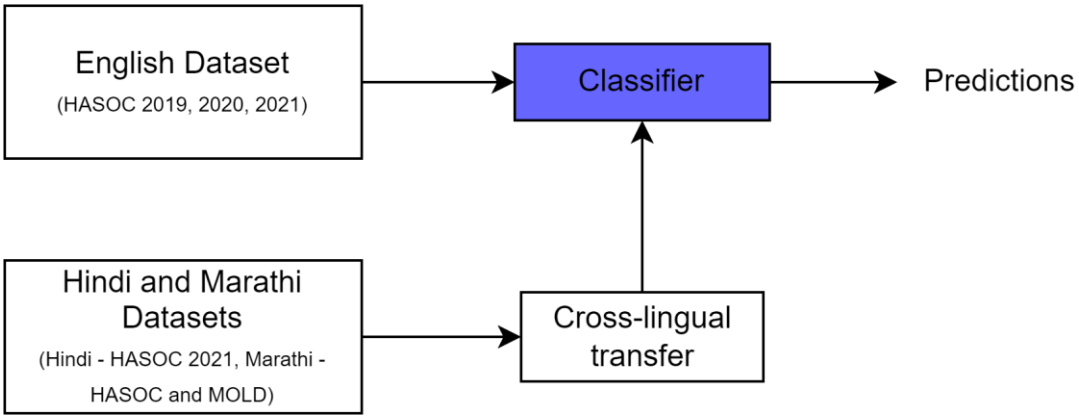
WORKFLOW

The workflow of the project involves working on the following steps:-

- Data Collection and Description
- Selecting Text Classification Approaches
- Fine tuning models using English, hindi and marathi hate speech data
- Model Training and Hyperparameter Tuning
- Evaluation Metrics and Results



Flow of binary classification



MILESTONES OVERVIEW

MANDATE 2: Lexical Processing

- Data acquisition for English, Hindi, Marathi languages.
- Text Cleaning
- Label Encoding
- Stopwords removal
- Tokenization
- Lemmatization, segmentation
- Embeddings

MANDATE 3: Syntax Processing and Model Development

- POS Tagging
- Fine-Tuning Pre-trained Models on English Dataset: Pre-trained transformer based models (e.g mBERT) are fine-tuned on the hate speech detection task for better performance on the specific context.
- BERT models, including indicBERT, mBERT, RoBERTa, are employed for comparison.
- Learning rates are adjusted based on model performance during training.
- Dropout rates are tuned to prevent overfitting.
- Batch sizes and epochs are optimized for efficient model training.

MANDATE 4: Semantic Processing and Cross-lingual transfer

- Cross-lingual transfer, where a high-resource transfer language (English) is used to improve the accuracy of a low-resource task language (Hindi and Marathi)
- Fine-tuning a pre-trained model on a dataset containing multiple languages like English, Hindi, and Marathi aims to leverage the knowledge gained from one language to improve performance in other languages.
- Apply a classification layer on top of the model to make predictions
- Evaluation Metrics: Models are evaluated using various metrics, including accuracy, macro F1 score, precision, and recall.

FUTURE SCOPE

- Multilingual Expansion: Extend models beyond Hindi and Marathi for broader coverage across Indian languages. Exploring on code mix data.
- Fine-Grained Refinement: Improve fine-grained classification for Hindi, considering additional categories and optimizing hierarchical approaches.
- User Interface Integration: Explore integrating hate speech detection models into social media platforms via a user-friendly API for real-time content moderation.