P7165(CSE-DS)

Offensive Language Identification In Code-Mixed Text

Title: Detecting Offensive Language in Code-Mixed Text using N-gram Features and BERT Embedding

Abstract:

Objective: This project aims to develop a robust offensive language detection system using natural language processing (NLP) techniques, specifically tailored for code-mixed Kannada-English Language . Our goal is to automatically identify offensive content within the nuanced context of mixed-language usage.

Methods: Leveraging NLP methodologies, we employ a combination of Bert and word embeddings to effectively detect offensive language in code-mixed Kannada-English Comments. We explore transfer learning and linguistic preprocessing techniques to enhance model adaptability.

Dataset: The project capitalizes on a meticulously curated dataset containing code-mixed Kannada-English Youtube Comments, ensuring the availability of contextualized data for training and evaluating the offensive language detection models.

Significance: Addressing the imperative need for accurate offensive language identification in social media, this project contributes to fostering respectful online interactions and content moderation, particularly in the unique context of mixed-language content.

Methodology: Our NLP-centric approach encompasses linguistic preprocessing, innovative model architecture design sensitive to code-mixed text, model training, and evaluation based on F1-score and precision metrics.

Innovation: We introduce an attention-based mechanism tailored to capture the intricacies of offensive language nuances present in code-mixed Kannada-English Language.

Scope: The project confines its scope to the identification of offensive language within code-mixed Kannada-English text in Youtube comments, focusing solely on detection rather than sentiment analysis or broader user interactions.

Conclusion: This endeavor establishes the groundwork for an effective NLP-driven offensive language detection system, catering specifically to code-mixed Kannada-English Youtube Comments on Youtube. By harnessing the capabilities of NLP and deep learning, this project showcases potential for nurturing a positive and inclusive online discourse.

Keywords: Offensive language detection, NLP, recurrent neural networks, attention mechanisms, code-mixed Kannada-English text, social media content moderation, linguistic preprocessing, word embeddings.

Introduction:

In today's digital age, social media platforms have become a prominent space for communication, expression, and interaction. However, this widespread online presence has brought to light challenges related to the quality of interactions and content shared. Offensive language and inappropriate content have become persistent concerns, leading to the need for robust systems that can automatically detect and mitigate such instances. This project embarks on the journey of developing a sophisticated offensive language detection system, specifically tailored for code-mixed Kannada-English text in Youtube-Comments on Youtube, using Natural Language Processing (NLP) techniques.

The Context of Code-Mixed Language:

In multilingual societies, individuals often communicate in code-mixed language, which involves seamlessly blending two or more languages within a single conversation or text. This linguistic phenomenon poses a unique challenge for existing offensive language detection systems, as the contextual nuances and interplay between languages can impact the accuracy of detection. The Kannada-English language, rich in heritage and widely spoken, often intertwines with other languages in online conversations. Therefore, this project focuses on developing an offensive language detection system that can navigate the intricacies of code-mixed Kannada-English text on the Youtube platform.

The Role of Natural Language Processing (NLP):
Natural Language Processing has emerged as a vital field within Artificial Intelligence, enabling machines to comprehend, analyze, and generate human language. Leveraging NLP techniques, this project aims to build an advanced model capable of understanding the complex linguistic patterns, cultural contexts, and linguistic variations present in code-mixed Kannada-English Language. By harnessing the power of NLP, we seek to develop a system that can effectively distinguish offensive language from legitimate expressions, fostering a more inclusive and respectful online environment.

Project Objectives and Methodologies:
The primary objective of this project is to create a state-of-the-art offensive language detection system for code-mixed Kannada-English text on Youtube. To achieve this, we employ a combination of advanced NLP techniques and deep learning methodologies. Our approach encompasses the utilization of recurrent neural networks (RNNs) and attention mechanisms, which are specifically designed to capture the intricacies of language mixing. Moreover, we explore the integration of transfer learning and linguistic preprocessing techniques to enhance the model's adaptability and accuracy.

Significance and Potential Impact:
The significance of this project lies in its potential to address a critical challenge in the realm of online communication. By developing an accurate and effective offensive language detection system, we aim to contribute to the mitigation of harmful content and offensive language, fostering a safer and more respectful online space. This system has the potential to be adopted by social media platforms, content moderation teams, and online communities, enabling proactive management of inappropriate content and safeguarding user experiences.

Related Works:

The landscape of offensive language detection has primarily revolved around monolingual text, often overlooking the intricate challenges posed by code-mixed content. Early research in the field employed rule-based approaches and lexicon-based methods to identify offensive language based on predefined lists of offensive words and patterns. However, these methods suffer from limited coverage and struggle to account for context-dependent variations in offensive language usage.

Recent advancements have witnessed the emergence of machine learning techniques, especially deep learning models, in the domain of offensive language detection. Notably, models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated considerable success in learning contextual patterns and features that indicate offensive content. While these models have shown promise, they often rely on large annotated datasets, which might not be readily available for all languages and language combinations.

The introduction of transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers), revolutionized the field by capturing contextual information more effectively. BERT's bidirectional attention mechanism allows it to understand the context of a word within a sentence, leading to better embeddings that encapsulate the meaning and nuances of the text. This capability has proven crucial in understanding the sentiment, intent, and offensive nature of text.

However, the majority of offensive language detection research has been confined to monolingual contexts, predominantly in English. The transition to code-mixed content adds complexity due to the blending of multiple languages within a single text. While some efforts have been made to apply existing models to code-mixed text, the results have often been suboptimal due to the unique linguistic characteristics, context shifts, and language mixing patterns present in such content.

Our work bridges this gap by extending offensive language detection to code-mixed text in Dravidian languages, including Tamil, Malayalam, and Kannada. We build upon the insights gained from previous research in monolingual and multilingual contexts, incorporating the strengths of BERT embeddings to capture contextual information. Additionally, we leverage n-gram features to capture local patterns that might signal offensive language, addressing the specific intricacies of offensive language identification in code-mixed contexts. Through our novel approach, we aim to tackle the challenges associated with identifying offensive language in Dravidian languages and offer insights that contribute to the broader field of multilingual offensive language detection.

Methodology:

Our proposed methodology for detecting offensive language in code-mixed text combines the power of BERT embeddings and n-gram features. This approach is designed to tackle the challenges posed by offensive language identification in code-mixed Dravidian languages, considering their unique linguistic characteristics and the intricacies of multilingual content.

- **BERT Embeddings:**

We leverage BERT, a state-of-the-art transformer-based model, to generate contextual word embeddings. BERT captures the semantic and contextual information of words by considering their surrounding words bidirectionally. This allows the model to understand the meaning and nuances of words within a sentence. In our approach, we utilize pre-trained BERT models fine-tuned on multilingual data to generate embeddings for the words in the code-mixed text. These embeddings encode both the syntactic structure and the semantic context of the text, enabling us to capture the complexity of offensive language in multilingual content.

- **N-gram Features:**

Recognizing the importance of local patterns and specific word combinations that signify offensive language, we incorporate n-gram features into our approach. N-grams are contiguous sequences of n items from a given sample of text, where n represents the number of words or characters in the sequence. By extracting n-grams from the code-mixed text, we aim to capture both common and context-specific word patterns that might indicate offensive content. This allows our model to consider not only the broader context provided by BERT embeddings but also the finer linguistic details present in the n-gram features.

- **Feature Fusion and Classification:**

To effectively combine the information from BERT embeddings and n-gram features, we employ a feature fusion mechanism. This mechanism enables the model to weigh the importance of different features based on their relevance to offensive language identification. The fused features are then fed into a classification model, which learns to distinguish between offensive and non-offensive language patterns. We opt for machine learning algorithms known for their performance in text classification tasks, such as Support Vector Machines (SVM) or neural network-based classifiers.

- **Training and Evaluation:**

Our approach is trained on a carefully curated dataset of code-mixed text in Kannada-English language, which includes labeled instances of both offensive and non-offensive language. During training, the model learns to recognize the contextual cues and patterns indicative of offensive language, optimizing its parameters using labeled data. We evaluate the model's performance using metrics like precision, recall, and F1-score on a separate test dataset, ensuring that it generalizes well to unseen instances of code-mixed text.

By combining the strengths of BERT embeddings and n-gram features, our methodology addresses the intricacies of offensive language detection in code-mixed Kannada-English languages. This hybrid approach allows us to capture both global context and local linguistic patterns, providing a comprehensive understanding of offensive language within multilingual content. Through our methodology, we strive to achieve state-of-the-art results in detecting offensive language in code-mixed text, thereby contributing to advancements in multilingual natural language processing and offensive content moderation.

Conclusion:

This robust offensive language detection approach tailored for code-mixed text in Kannada-English languages. By amalgamating n-gram features and BERT embeddings, our method achieves state-of-the-art results on a new dataset of YouTube comments. This work contributes to the advancement of offensive language detection in multilingual and code-mixed contexts, and opens avenues for further research in adapting and refining these techniques for other languages and domains.