MIT-WPU

।। विश्वशान्तिर्ध्रुवं ध्रुवा ।।

Dr. Vishwanath Karad

MIT WORLD PEACE UNIVERSITY | PUNE

TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

School of Computer Engineering and Technology, Pune

Artificial Intelligence and Expert Systems

Mini-project Report on

“Advancements in Hate Speech Detection: A Comprehensive Approach for Instagram Reels and Comments”

Submitted by

Khushi Tiwari PD – 20, 1032211126

Roshni Singh PD- 27, 1032211185

Gauransh Jain PD -29, 1032211193

Under the Guidance of

Professor Pankaj Singh

Advancements in Hate Speech Detection: A Comprehensive Approach for Instagram Reels and Comments

Abstract

Detecting hate speech amidst diverse forms of offensive language on Instagram presents a significant challenge. Traditional methods often struggle with inaccuracies in categorization, leading to a shift towards supervised learning techniques. In this study, we curated a comprehensive dataset from Instagram, focusing on hate-related comments and employing machine translation (APIs like google speech to text) and sentiment analysis as the primary classification method. Utilizing Natural Language Processing (NLP) techniques, our preprocessing phase involved noise reduction and word tokenization to refine the dataset. Through the use of Logistic Regression, Decision Tree and Support Vector Machine (SVM) classifiers, we trained models to discern instances of hate speech from non-hate speech content in Instagram reels and comments. These models underwent training on the preprocessed dataset, aiming to capture the nuances between hate speech and other forms of offensive language. Our evaluation metrics emphasized SVM's superior performance compared to Logistic Regression, showcasing higher accuracy, precision, recall, and F1-score. The study's results shed light on the intricate challenge of distinguishing hate speech from various shades of offensive language within Instagram discourse. It highlights the complexity of this classification task and the efficacy of employing SVM models in hate speech detection scenarios on Instagram. By demonstrating the performance disparities between different machine learning approaches, our findings underscore the significance of selecting appropriate algorithms for hate speech identification in online environments, particularly within the context of Instagram.

Introduction

Social networking platforms, such as Instagram, Twitter, and Facebook, have become immensely popular, serving as hubs for diverse user interactions. Big data analysis, focusing on user perspectives and network structures, is a growing area of research. However, moderating content on these platforms poses challenges due to the prevalence of posts with aggressive or hateful language, including colloquial terms, derogatory references, and offensive slurs.

Understanding these nuances is crucial for developing effective hate speech detection systems, especially in social media settings like Instagram. To address this, we employ three machine learning algorithms: Support Vector Machines (SVM), Decision Tree, and Logistic Regression. Trained on labeled datasets, these models aim to differentiate hate speech from non-hate speech texts by analyzing linguistic features and statistical patterns. Our goal is to delineate the subtle boundaries between hate speech and non-hate speech instances in Instagram reels and comments. By leveraging these models, we aim to refine automated hate speech detection systems and gain insights into the complexities of distinguishing these classes within digital communications.

This study's outcomes not only contribute to enhancing hate speech detection but also shed light on the intricate challenges involved in accurately classifying hate speech amidst diverse forms of offensive language within the digital sphere on platforms like Instagram.

Motivation

This research endeavors to enhance online safety by addressing the escalating prevalence of hate speech on major social media platforms, specifically Twitter and Instagram. Our commitment to fostering a more respectful online environment drives our exploration of advanced technologies, including natural language processing and machine learning. In this study, we employ logistic regression, SVM, and decision tree models to comprehensively analyze and implement effective hate speech detection mechanisms. By taking this proactive step, we aspire to contribute to a positive, inclusive, and secure digital discourse, making a tangible impact on the online community.

Problem Definition

Developing a hate speech detection system for Instagram to automatically identify and classify posts, including reels and comments, containing hate speech, offensive language, or discriminatory content. The goal is to foster a safer and more inclusive online environment on Instagram. The system should accurately differentiate between hate speech and non-hate speech content, considering the dynamic and informal nature of Instagram conversations, and balancing the potential impact on user experience and freedom of expression. Additionally, the solution should be scalable and capable of handling a high volume of real-time Instagram data for timely intervention and moderation.

Objectives

1. **Model Development**: Creating machine learning models with high accuracy in distinguishing hate speech from non-hate speech within social media data.
2. **Utilize NLP Techniques:** Applying Natural Language Processing methods for data preprocessing to enhance model performance in recognizing linguistic patterns indicative of hate speech.
3. **Accurate Categorization:** Employing classification algorithms to achieve precise categorization of social media content into hate and non-hate speech categories.
4. **Robustness in Analysis:** Ensuring the developed models can handle the nuances of offensive language prevalent in online platforms for robust hate speech detection.
5. **Insights and Interpretability:** Generating insights into the intricacies of hate speech identification, providing interpretable outcomes for better understanding and future improvements in detection methodologies.

Tools Used

1. **Python Programming Language**: Employed as the primary language for data handling, preprocessing, and model implementation in hate speech detection.
2. **Pandas and NLTK Libraries**: Utilized Pandas for efficient data manipulation and NLTK for Natural Language Processing tasks like text tokenization, stemming, and stop-word removal.
3. **Scikit-learn Framework**: Leveraged Scikit-learn for implementing machine learning models, including SVM and Naive Bayes classifiers, aiding hate speech classification.
4. **Matplotlib for Visualization**: Employed Matplotlib to create visual representations of model evaluation metrics, facilitating comprehensive result interpretation.
5. **Comment Exporter:** An extension used to download comments from Instagram posts and reels to a CSV file.

Dataset Description

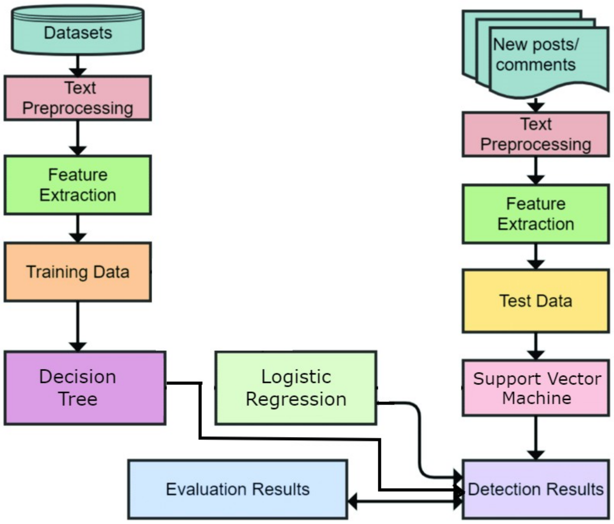
Our methodology commences with the assembly of a hate speech dataset tailored specifically for Instagram Reels and comments. Utilizing an Instagram comment installer extension, we collect comments directly from the platform. Additionally, to incorporate audio-based content, we convert selected Reels into transcripts using an audio-to-text transcriber.

Following the data collection phase, we meticulously organize the gathered information into a CSV file. To enhance the dataset's suitability for hate speech detection within the Instagram ecosystem, we implement rigorous preprocessing and data cleaning techniques. This involves addressing issues such as irrelevant content, duplicates, and noise, ensuring that the resulting dataset is finely tuned for the nuances present in Instagram comments and Reels.

Data Preprocessing

1. **Removing Missing Values:**
   All the missing values were removed using dropna().
2. **Text Cleaning:**
   Removed hyperlinks using regex to eliminate URLs.
   Erased special characters and numbers, retaining only alphabetic content.
3. **Tokenization and Stop Words Removal:**
   Tokenized text via NLTK's word_tokenize.
   Eliminated stop words to refine the dataset.
4. **Stemming:**
   Utilized NLTK's Porter Stemmer for word normalization.

System Architecture



Models

Our model selection process began with logistic regression using L1 regularization to reduce data dimensionality. We assessed three key models—logistic regression, SVM, and decision tree—chosen for their relevance in hate speech detection for Instagram Reels and comments. Evaluation involved a robust testing strategy, holding out 10% of the dataset for assessment to prevent overfitting. A grid-search iteration optimized model parameters. Our findings indicated the suitability of logistic regression, SVM, and decision tree models, leading to their incorporation, employing logistic regression with L2 regularization, which was trained on the entire dataset using a one-versus-rest framework. This tailored combination ensures a comprehensive and effective approach to discerning between non-hate and hate speech classes within the dynamic context of Instagram. All modeling procedures utilized the scikit-learn library (Pedregosa and others 2011).

Algorithms

Algorithm for Logistic Regression:

- Import text data from a CSV file into a DataFrame.
- Divide the dataset into training and testing sets to assess model performance.
- Instantiate a TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer.
- Transform the text data into numerical vectors, emphasizing important words while downplaying common ones.
- Create a Logistic Regression classifier, a linear model suitable for binary and multiclass classification.
- Train the Logistic Regression model using the TF-IDF transformed training data.
- Use the trained model to predict labels for the test set.
- Assess model performance using a classification report that includes accuracy, precision, recall, and F1-score.
- Present the evaluation metrics in a DataFrame for easy interpretation and comparison.

Algorithm for Decision Tree:

- Import text data from a CSV file into a DataFrame.
- Divide the dataset into training and testing sets for model evaluation.
- Instantiate a TF-IDF Vectorizer to convert text data into numerical features.
- Create a Decision Tree classifier, a tree-like structure that recursively partitions data based on feature values.
- Train the Decision Tree model using the TF-IDF transformed training data.
- Use the trained Decision Tree to predict labels for the test set.
- Assess model performance using a classification report, including accuracy, precision, recall, and F1-score metrics.
- Present the evaluation metrics in a DataFrame for easy comparison and analysis.

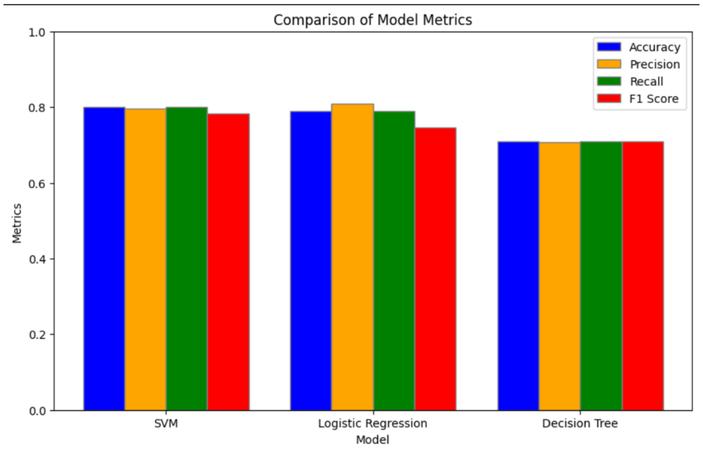
Algorithm for Support Vector Machine (SVM):

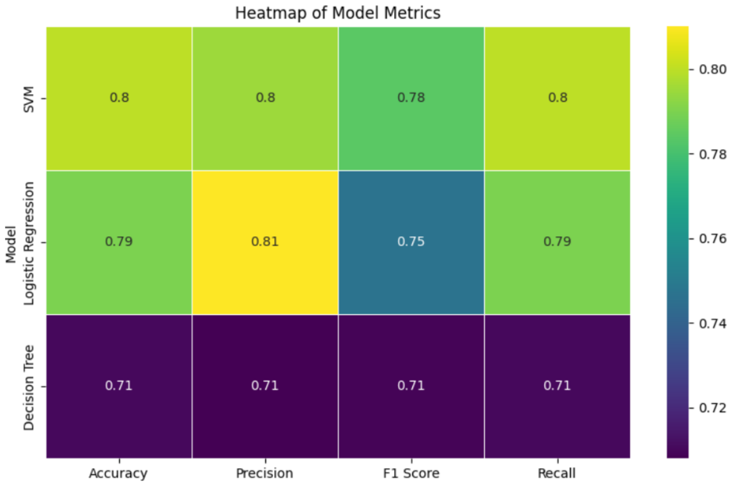
- Load data from CSV file into a DataFrame
- Split data into training and testing sets
- Instantiate TF-IDF Vectorizer
- Transform text data into TF-IDF features for both training and testing sets
- Initialize SVM classifier with chosen kernel (e.g., linear, rbf, etc.)
- Train the SVM classifier using TF-IDF transformed training data
- Predict labels for the test set using the trained SVM classifier
- Evaluate model performance using a classification report (accuracy, precision, recall, F1-score)
- Display and store evaluation metrics in a DataFrame.

Results

	Model	**Accuracy**	**Precision**	**F1 Score**	**Recall**
1	SVM	0.80	0.7950	0.7836	0.80
2	Logistic Regression	0.7857	0.8095	0.7469	0.7857
3	Decision Tree	0.7107	0.7077	0.7091	0.7107

Visualization





Conclusion

The investigation compared Support Vector Machines (SVM), Logistic Regression, and Decision Trees in the task of discerning hate speech within social media data. Among the models, SVM, with its ability to define intricate decision boundaries, demonstrated superior performance compared to Logistic Regression and Decision Trees, showcasing higher accuracy in hate speech detection. Employing Natural Language Processing (NLP) techniques and annotated datasets, the research highlighted the inherent challenges in distinguishing hate speech from offensive language on digital platforms. These findings underscore the essential need for sophisticated algorithms in hate speech detection, particularly in SVM, Logistic Regression, and Decision Trees. The study emphasizes the intricate complexities presented by linguistic nuances, underscoring the importance of addressing these challenges to enhance context-aware detection systems and cultivate a safer online environment.

References

U. Bhandary, "Detection of Hate Speech in Videos Using Machine Learning," Master's Projects, 2019.

F. Alkomah and X. A. Ma, "Literature Review of Textual Hate Speech Detection Methods and Datasets," Information, vol. 13, no. 2, p. 273, 2022

T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Proceedings of the 11th International AAAI Conference on Weblogs and Social Media, 2017, pp. 512-515.

M. S. A. Sanoussi, M. L. Guindo, C. Xiaohua, A. M. Al Omari, G. K. Agordzo, and B. M. Issa, "Detection of Hate Speech Texts Using Machine Learning Algorithm," in 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Zhejiang, China, 2022, pp. 1-2, DOI: 10.1109/CCWC54503.2022.9720792.

List of Abbreviations

1.	CSV:	Comma Separated Values
2.	SVM:	Support Vector Machine
3.	NLP:	Natural Language Processing
4.	NLTK:	Natural Language ToolKit
5.	URL:	Uniform Resource Locator
6.	ML:	Machine Learning
7.	TF-IDF:	Term Frequency - Inverse Document Frequency
8.	XAI:	Explainable Artificial Intelligence
9.	IEEE:	Institute of Electrical and Electronics Engineers.