SHIELD: Social Hate Identification and Emotion Led Detection

[1] Mathuvanthi V, [2]Nivethidha S, [3] Ramya S, [4] Mrs. M. Kasthuri
[1]Student, [2]Student, [3]Student, [4] Asst. Professor
[1]Computer Science and Engineering, [1]Adhiyamaan College Of Engineering, Hosur, India

Abstract: **Social media allows users to express themselves freely, but it also facilitates the spread of hate speech, particularly in comment sections. This project, titled " Social Hate Identification And Emotion Led Detection " proposes a Java-based software framework for detecting and analyzing hate speech, with a focus on fat-shaming content. The framework utilizes sentiment and emotion analysis to identify and categorize hate speech, converting unstructured data into structured formats for further study. By integrating sentence classification techniques, it enhances detection accuracy. This project aims to support researchers and policymakers by providing an automated tool to analyze and combat online hate speech effectively.**

Keywords: **Hate Speech Detection, Social Media Analysis, Sentiment Analysis, Emotion Analysis, Fat-Shaming, Unstructured Data, Data Integration, Sentence Classification, Java Framework, Online Hate Speech.**

INTRODUCTION

Hate speech is a significant problem on social media platforms, with potentially severe consequences for individuals and society as a whole. The rise of social media has facilitated the spread of hateful and discriminatory messages, which can fuel prejudice, bigotry, and even violence.To address this issue, we propose a framework called SHIELD, which stands for " Social Hate Identification And Emotion Led Detection " SHIELD aims to detect hate speech on Facebook pages and integrate unstructured data through sentiment and emotion analysis.

The proposed framework consists of four stages: data collection and preprocessing, sentiment and emotion analysis, hate speech clustering, and evaluation. In the first stage, the system collects the data from Facebook pages and preprocesses it to remove noise and irrelevant information. In the second stage, the system performs sentiment and emotion analysis to understand the underlying emotions and sentiments of the text. In the third stage, the system clusters the text based on the degree of hate speech expressed. Finally, in the fourth stage, the system evaluates the effectiveness of different strategies for data analysis and natural language processing. The system provides two types of applications, one that uses a dataset and one that allows users to post comments. The admin part of the application allows for monitoring of sentiment analysis and other results, which can help to identify potential issues early and take appropriate actions to address them.

Overall, the SHIELD system offers a powerful tool for detecting and addressing hate speech on social media platforms. By identifying the most significant factors from the unstructured data, the system can effectively cluster information into different groups according to the degree of hatred being expressed. The system has the potential to make a significant contribution towards creating a more respectful and inclusive online community.

Recent Works

Several studies have explored SHIELD to detect hate speech on Facebook pages and integrate unstructured data through sentiment and emotion analysis to enhance the experience. The following research works provide insights into various aspects of this domain:

Matamoros-Fernández & Farkas (2019) [1] conducted a systematic review on racism and hate speech in social media research, building upon Jessie Daniels's 2013 work. Examining 104 articles, they analyzed the geographical contexts, platforms, and methodologies used in studying online racism. Their findings highlight a lack of diversity in geographic and platform representation, limited reflexivity from researchers regarding their subjects, and minimal engagement with critical race perspectives. The study

underscores the need for more thorough investigations into how user behaviors and platform policies shape contemporary racism.

Del Vigna et al. (2017) [2] explored hate speech detection on Facebook, focusing on textual content in Italian public pages. They proposed a classification taxonomy for different hate categories and annotated comments using multiple human evaluators. Utilizing Support Vector Machines (SVM) and Long Short-Term Memory (LSTM) networks, they developed two classifiers for hate speech recognition in Italian. Their experiments demonstrated the effectiveness of these machine learning approaches, establishing the first manually annotated Italian Hate Speech Corpus for social media text analysis.

Ahmed et al. (2020) [3] provided a comprehensive survey on the k-means clustering algorithm, addressing its limitations such as sensitivity to centroid initialization and predefined cluster numbers. The paper discussed various enhancements to k-means, including algorithmic modifications to improve robustness and adaptability to different data types. Through extensive experimental analysis, the study compared multiple k-means variants and provided insights into their effectiveness, contributing to a deeper understanding of clustering methodologies.

Santia & Williams (2018) [4] introduced BuzzFace, a dataset designed for news veracity assessment on Facebook. The dataset, derived from nine news outlets' Facebook posts and annotated by BuzzFeed, categorizes articles into four veracity levels: mostly true, mostly false, mixed, and no factual content. Integrating Facebook comments, reactions, and embedded tweets, BuzzFace enables advanced machine learning applications for fake news detection and bot identification. With over 1.6 million text entries, the dataset significantly surpasses existing gold-standard datasets in scale and complexity.

Franzoni et al. (2019) [5] developed a path-based emotion abstraction model for analyzing Facebook comments through sentiment analysis and taxonomy knowledge. Recognizing the role of biases in shaping user engagement, the study introduced an automated clustering approach to extract emotional sub-contexts. Their methodology, validated through clustering techniques and expert evaluation, demonstrated alignment with human perception in grouping Facebook comments based on emotional valence. The research contributes to the understanding of online emotional expression and its influence on social media discourse.

Existing System

The authors of "Hate Me, Hate Me Not: Hate Speech on Facebook" have proposed several classification methods to distinguish among different types of hate speech. More specifically, they leverage morpho-syntactic features, sentiment polarity, and word-embedded lexicons to design and implement two classifiers for Italian. Their framework utilizes support vector machines (SVMs) and long short-term memory (LSTM) networks. This study was premised on the concept outlined in Del Vigna et al.'s study and our understanding of hate speech. Another major challenge in traditional retail environments is inefficient inventory management. Many supermarkets rely on periodic manual stock checks, which are time-consuming and prone to errors. Stock levels are often updated with delays, leading to discrepancies between actual inventory and recorded data. This inaccuracy causes stock outs, which frustrate customers and result in lost sales, or overstocking, which ties up capital in unsold goods. Additionally, without real-time tracking, retailers struggle to predict demand accurately, making it difficult to optimize stock replenishment and promotional strategies. Shrinkage due to theft, misplacement, or human error further exacerbates inventory management issues, leading to financial losses.

In the existing system first they recognize a set of pages from American-based websites, known to discuss controversial topics such as immigration, race, and religion. We use these seeds, or Facebook IDs, to crawl the Facebook graphs and construct a small network using the follow relationship. Leveraging graph analysis techniques, we identify the most influential pages spreading hate speech and crawl their latest posts and comments. Some supermarkets have attempted to address these challenges by introducing self-checkout kiosks, but these systems come with their own limitations. The high initial investment required for self-checkout hardware and maintenance makes it less feasible for small and mid-sized retailers. Additionally, some customers find self-checkout systems difficult to use,

particularly elderly shoppers who may struggle with technology. Fraudulent activities such as incorrect barcode scanning, item swapping, or bypassing payment are common issues in self-checkout systems, requiring constant monitoring and intervention from store employees.

We then apply sentiment and emotion analysis algorithms to recognize posts with highly negative tones, specifically those suspected of instigating hatred. Finally, we convert each post into a document by concatenating all comments, and using the K-means algorithm to create clusters of posts based on the topics they discuss.

Proposed System

In this proposed system we aim to create a framework (SHIELD) that can first detect hate speech on Facebook and then integrate unstructured data through clustering using sentiment and emotion analysis. It identifies the most significant factors from the unstructured data of posts and comments on Facebook pages that allegedly promote hate speech. We will conduct experiments to measure the effectiveness of different strategies for data analysis and natural language processing by implementing SHIELD.

Ultimately, we can demonstrate the strong effectiveness of hate-speech clustering using hybrid methods of data analysis and natural language processing to identify and categorize information into different groups according to the degree of hatred being expressed.

In Our Proposed System we develop 2 types of application where the first one use dataset and the second one using comments posted by the user. In the first type the facebook comments which are referred from kaggle website, are uploaded into the system. Then the preprocessing is done and each and every comment given in the dataset is processed one by one and the Sentiment Score is calculated and the Sentiment type is predicted for the each dataset record.

Additionally it computes the Emotion Score, Emotion Type (Happiness, Sadness, Anger, Fear, Disgust, Surprise), Hate Word and Label the dataset record completely. Finally a static Graph is plotted with the results which we received.

In the second type we develop an application with two entities: Admin and user. Where user can post the comments and in the admin part the Sentiment Analysis and other results will be monitored. In the admin part mainly we predict the Sentiment Score, Sentiment type, Emotion score, Emotion type and Hate Word (Example: Fuck, assassin, stupid fucker, criminal, negro, nasty bitch, lesbian, israel's, bullshit, homosexual etc.)

METHODOLOGY

The SHIELD system follows a well-structured methodology to effectively detect and analyze hate speech on Facebook using Natural Language Processing (NLP) techniques.

The application is developed using Java on NetBeans 8.2, ensuring a stable and scalable environment for efficient data processing. The MySQL database is used for structured storage of Facebook comments and hate speech-related data, enabling fast retrieval and analysis. The system operates on a Windows 10 platform, utilizing a Pentium i3 processor with 2GB RAM and a 500GB hard disk for smooth execution.

The SHIELD framework consists of two main applications:

1. Dataset-Based Hate Speech Detection: This module processes Facebook comments collected from Kaggle datasets, performing preprocessing, sentiment analysis, and emotion detection. Each comment undergoes sentiment scoring and classification into emotions like happiness, sadness, anger, fear, disgust, and surprise. The system also identifies hate words and labels comments accordingly. Results are visualized through static graphical representations for better insights.
2. Real-Time Comment Analysis: This module enables user-admin interaction, where users post comments, and admins monitor sentiment and hate speech detection in real time. The admin dashboard provides insights into sentiment scores, emotion analysis, and hate word detection, allowing manual review and moderation.

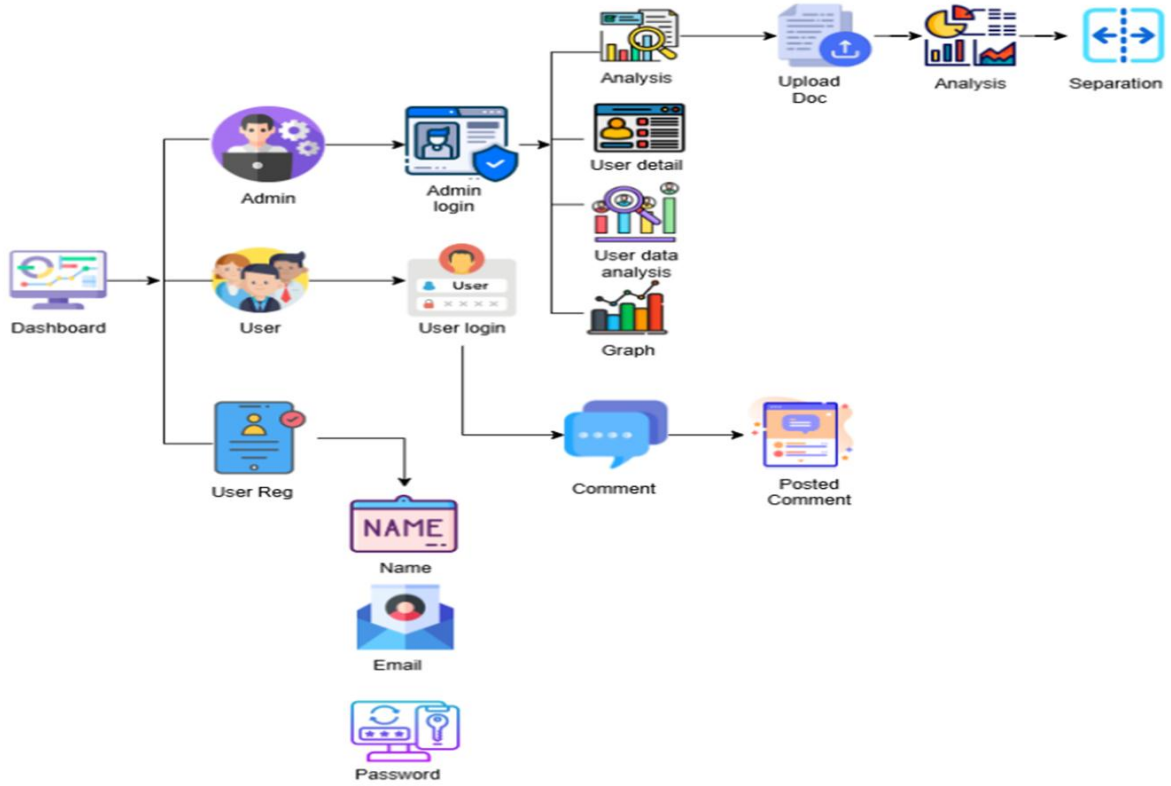
The Natural Language Processing (NLP) techniques used in the system ensure high accuracy in text classification and clustering. Hybrid clustering methods categorize comments based on severity of hate speech, improving detection efficiency. The

MySQL database ensures structured data management, allowing efficient storage and retrieval of results.

The system is designed to be scalable and adaptable, with potential for multi-platform integration in future versions. Graphical visualizations enhance reporting, while automated analysis improves efficiency in hate speech detection. This structured methodology ensures reliable, real-time, and data-driven social media monitoring, contributing to a safer digital environment.

Modeling

The System is modelled as given below.



RESULTS

- Efficient Hate Speech Detection: The proposed SHIELD framework effectively detects and categorizes hate speech from Facebook posts and comments.
- Comprehensive Sentiment & Emotion Analysis: The system analyzes comments for sentiment type, sentiment score, emotion score, and emotion type (e.g., happiness, sadness, anger, fear, disgust, surprise).
- Structured Hate Speech Labeling: Identifies and labels hateful words (e.g., "fuck," "assassin," "criminal," "bullshit") to classify the degree of hate speech.
- Data-Driven Hate Speech Clustering: Uses hybrid clustering methods to categorize hate speech based on severity, improving analysis accuracy.
- Automated Dataset Processing: The first application processes Kaggle-based Facebook comments, automating sentiment and emotion scoring for large datasets.
- Interactive User-Admin System: The second application allows real-time comment posting by users, with administrators monitoring and analyzing hate speech.
- Graphical Representation of Insights: The system generates static graphs to visualize sentiment scores, emotion types, and hate speech trends.
- Enhanced Moderation Capabilities: Provides admins with insights into potential hate speech trends, helping them take necessary actions.
- Scalable NLP Implementation: Leverages Natural Language Processing (NLP) to improve detection efficiency across different datasets.
- Improved Social Media Monitoring: Enables automated and manual monitoring of hate speech on social media platforms, ensuring a safer digital space.

Advantages

- Accurate Hate Speech Detection: Utilizes advanced NLP techniques to improve the accuracy of identifying hate speech in Facebook comments.
- Efficient Data Processing: Automates sentiment and emotion analysis, reducing manual effort and processing time for large datasets.
- Real-Time Hate Speech Monitoring: Provides instant analysis of user-posted comments, enabling proactive moderation.
- Enhanced Content Moderation: Helps admins identify and take action against harmful content more effectively.
- Comprehensive Sentiment & Emotion Analysis: Identifies not just hate speech but also underlying emotions like anger, sadness, or fear, offering deeper insights.
- Improved Social Media Safety: Reduces the spread of hate speech, promoting a healthier online environment.
- Data-Driven Decision Making: Generates insights based on sentiment scores, hate speech trends, and emotional impact, aiding in better policy-making.
- User-Admin Interaction System: Allows real-time user engagement with an admin monitoring system for enhanced moderation.
- Scalable & Flexible Implementation: Can be adapted to different social media platforms and datasets for wider applications.
- Graphical Visualization of Insights: Provides easy-to-understand visual reports, helping administrators and researchers analyze trends effectively.

FUTURE SCOPE

Future enhancements for the SHIELD Framework include:

- Enhanced Real-Time Monitoring: Future versions can integrate AI-driven real-time monitoring tools to detect and block hate speech as soon as it is posted.
- Multilingual Hate Speech Detection: Expanding the system to support multiple languages will make it more effective for global social media platforms.
- Context-Aware NLP Models: Improving NLP models to understand the context of words and phrases will enhance accuracy in hate speech detection.
- Automated Moderation & Response: Implementing AI-based auto-response mechanisms to warn or block users posting hateful content.
- Integration with Other Social Media Platforms: Extending the framework beyond Facebook to platforms like Twitter, Instagram, and YouTube for a broader impact.
- Deep Learning-Based Hate Speech Categorization: Utilizing advanced deep learning techniques such as transformers for more precise hate speech categorization.
- User Behavior Analysis: Analyzing patterns in user behavior to predict potential hate speech before it escalates.
- Dynamic Graphical Insights: Enhancing visualization tools with interactive dashboards for better hate speech trend analysis.
- Collaboration with Law Enforcement & Social Platforms: Assisting regulatory authorities in identifying and mitigating harmful online activities.
- Ethical AI Implementation: Developing fair and unbiased AI models that minimize false positives and negatives in hate speech detection.

CONCLUSION

In conclusion, the proposed system, SHIELD, offers a comprehensive framework for detecting hate speech on Facebook and integrating unstructured data through sentiment and emotion analysis. By identifying the most significant factors from the unstructured data, the system can effectively cluster information into different groups according to the degree of hatred being expressed. This can help to prioritize and target resources towards the most severe cases of hate speech.

The system provides two types of applications, one that uses a dataset and one that allows users to post comments, which ensures that the system can be used in a variety of contexts and can be adapted to different user needs. The admin part of the application allows for monitoring of sentiment analysis and other results, which can help to identify potential issues early and take appropriate actions to address them.

Overall, the SHIELD system offers a powerful tool for detecting and addressing hate speech on social media platforms. The system has the potential to make a significant contribution towards creating a more respectful and inclusive online community.

REFERENCES

[1] Matamoros-Fernández, A., & Farkas, J. (2019). Racism, hate speech, and social media: A systematic review of research trends. *Social Media + Society, 5*(3), 1-13.

[2] Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate speech detection on Facebook: A study on the Italian language. *CEUR Workshop Proceedings, 1983*, 86-95.

[3] Ahmed, S., Karim, A., & Shah, A. (2020). A comprehensive review on k-means clustering: Algorithmic modifications and applications. *Journal of Big Data, 7*(1), 1-35.

[4] Santia, G., & Williams, J. R. (2018). BuzzFace: A news veracity dataset with Facebook data. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1), 531-540.

[5] Franzoni, V., Riccardi, A., & Tamburini, F. (2019). A path-based emotion abstraction model for Facebook comment analysis. *Expert Systems with Applications, 127*, 264-277.