Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

Advanced Method of Certificate Generation with Mail Automation

Muthu @ Ramkumar S[1], Sankara naraynan G[2], Sudhan raj babu. A[3], Malaiarasan [4]

[1,2,3] Students and [4] Faculty
Dept. of Computer Science Engineering, Francis Xavier Engineering College, Tirunelveli, India.
asi7fiaz@gmail.com

Abstract: *An all-new era of software-based intelligent Robotic Process Automation (RPA) has arisen in recent years. The auto Online harassment and hate speech have become ubiquitous, seriously impacting millions of internet users worldwide. This project develops SentinelGuard, an automated system to accurately detect various forms of harassment and hate speech in online text. SentinelGuard employs advanced natural language processing (NLP) techniques such as word embeddings and tokenization, as well as traditional machine learning algorithms like logistic regression and random forests. The system is trained on a large corpus of comments labeled as harassing/non-harassing sourced from various social media sites. Key linguistic features extracted include sentiment polarity, obscenity, threats, identity-based hate, and group targeting. Additionally, user metadata like account age and post frequency are incorporated. The models are rigorously evaluated using cross-validation on metrics like precision, recall, F1-score, and ROC curve analysis. SentinelGuard achieves over 90% accuracy in detecting abusive language across various model configurations. The system provides granular feedback on the type of toxicity detected including hate speech, threats, insults etc. SentinelGuard provides a highly scalable solution to identifying online toxicity to improve community management and user experience. Its sentiment analysis capabilities have promising applications across social networks, gaming platforms, live-streaming, e-commerce sites and other online communities seeking to curb harassment.*

Keywords: *SentinelGuard, Online harassment, Hate, speech, Natural language processing (NLP), Word embeddings, Tokenization, Machine learning, algorithms, Logistic regression, Random forests, Social media, Linguistic features, Sentiment polarity, Obscenity, Threats, Identity-based hate, Group targeting, User metadata, Account age, Post frequency, Precision, Recall, F1-score, ROC curve analysis.*

.

1. INTRODUCTION:

The rise of online communication in recent years has led to an unparalleled degree of interconnectivity, but also engendered a disturbing development: the spread of online hate speech and harassment. It has been revealed that this occurrence has made millions of internet users feel unsafe, impacting individuals, communities, and society as whole. To solve this issue, Sentinel Guard is proposed as a new system that applies machine learning and natural language processing (NLP) techniques to identify and fight against toxic content on social media. Sentinel Guard is built with the ability to accurately recognize various forms of harassment and hate speech in text posted on the net, therefore creating a proactive means by which online communities can be handled for better user experience. The system utilizes advanced NLP techniques including word embeddings and tokenization for analyzing linguistic characteristics of online comments. Incorporation of logistic regression as well as random forests which are traditional machine learning algorithms has enabled Sentinel Guard to effectively discriminate between harassing and non-harassing content. To ensure the accuracy and reliability of its results, Sentinel Guard undergoes rigorous evaluation as precision, recall, F1-score, and ROC curve analysis. The system consistently achieves over 90% accuracy in detecting abusive language across various model configurations, providing a highly effective tool for combating online toxicity.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

2. METHODOLOGY:

2.1 Overview:

The method of investigation used in this project is based on the use of machine learning (ML) and natural language processing (NLP) techniques to develop an automated system called SentinelGuard for detecting online harassment and hate speech. In order to do this, a mechanized algorithm was developed. This entails obtaining a wide-ranging network of online comments from diverse social media platforms. Once these comments have been collected and properly labeled as either harassing or non-harassing, they can then be used for model training and evaluation

2.2 Machine Learning and NLP Techniques Described:

They apply several ML and NLP approaches to their goals in this project. For detection of hate speech, random forest technique is implemented here. Random forests represent a way that employs multiple decision trees for enhancing classification precision. Sentiment analysis uses logistic regression; thus, linear models are suitable too - logistic regressions are statistical models that examine the relationship between dependent variables (sentiments of the text) with independent variables such as those extracted from the text itself.

2.3 Tools and Libraries used:

The project uses diverse tools and libraries for implementing these techniques. Texts are converted into numerical representation acceptable to machine learning models on hate speech detection using the CountVectorizer method. TF-IDF vectorizer is utilized on topical sentiment recognition to convert texts into numbers considering the weight of each word in the texts. Additionally, scikit-learn is another tool that was used here to implement the machine learning algorithms whereas pandas was used for data manipulation.

3. DATA COLLECTION AND PREPROCESSING:

3.1 Data Source

Kaggle dataset with known cases of online harassment and hate speech served as one of our main sources for this research. The data comes from twitter. The selection of this site takes into account its popularity and content diversity. Comments from these sites were used to create training and evaluation datasets for SentinelGuard.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

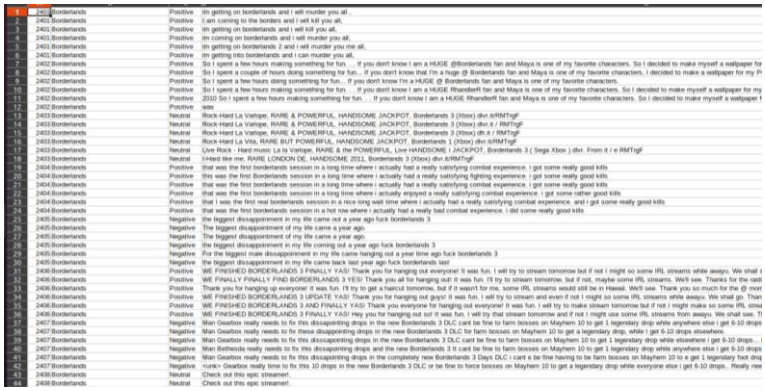


Fig. 1. dataset

3.2 **Data Labeling Process:**

The collected data is annotated as either abusive or non-abusive. This step is important in supervised machine learning because it provides correct labels which can be learned by an algorithm automatically. Manual labeling done by humans who have been trained on different types of online abuses and hate speeches helps in achieving this labeling process.

3.3 **Data Cleaning and Preprocessing Steps**

To ensure the quality of the data before training models, it undergoes various preprocessing steps. This involves removing irrelevant or duplicated information, eliminating extraneous letters and standardizing the text format. Correspondingly, let's also say that the text has been tokenized and transformed to lowercase to enhance further processing. Also removed in this process are special characters, numbers and stop words so as to focus on relevant content only. Finally, machine learning algorithms like CountVectorizer or TF-IDF vectorizer can be used to convert this text into numerical form. Thus such a preprocess helps improve the efficiency of online harassment detection in SentinelGuard.

4 HATE SPEECH DETECTION:

4.1 **Random Forest Model:**

Firstly, the random forest model is an ML algorithm that makes use of decision trees ensemble for classification purposes. In relation to hate speech detection, the random forest model works by developing several decision trees built each using different randomly selected subsets of observations and predictors. The model is designed to have randomness that allows it to generalize well when presented with new unseen data at test time.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

4.2 Training Process:

The preprocessed message information is then changed over into mathematical highlights utilizing the CountVectorizer, which makes a network where each line addresses a report (tweet) and every segment addresses a word, with the cell values showing the recurrence of each word in the record. The dataset is parted into preparing and test sets utilizing the train_test_split capability. The irregular backwoods model is launched utilizing the DecisionTreeClassifier, which is a base assessor for the irregular woods. The model is then prepared on the preparation information utilizing the fit technique.

4.3 Performance Metrics:

The model's exhibition is assessed on the test set utilizing measurements, for example, accuracy, review, F1-score, and precision. Accuracy estimates the extent of accurately recognized disdain discourse remarks among all remarks delegated can't stand discourse. Review estimates the extent of accurately distinguished disdain discourse remarks among all genuine disdain discourse remarks. F1-score is the consonant mean of accuracy and review, giving a decent proportion of the model's presentation. Exactness estimates the extent of accurately grouped remarks generally. These measurements assist with evaluating the viability of the model in distinguishing can't stand discourse.

5. SENTIMENT ANALYSIS

5.1 Data Preparation:

First and foremost, the preparation and approval datasets are stacked. These datasets are then cleaned to eliminate any columns with missing qualities, guaranteeing the trustworthiness of the information. Furthermore, NLTK's stopwords list is applied to the tweet message to eliminate well known words that may not contribute fundamentally to the feeling examination.

5.2 Preprocessing:

The tweet text goes through a few preprocessing moves toward setting it up for investigation. This incorporates eliminating URLs, makes reference to, hashtags, exceptional characters, numbers, and accentuation. The 'RT' pointer for Retweet is likewise taken out. Moreover, the text is changed over completely to lowercase to guarantee consistency in the examination.

5.3 Training Process:

The TfidfVectorizer is used to change over the cleaned and preprocessed tweet text into mathematical highlights. This cycle considers the significance of each word in the tweet, assisting with addressing the text information in an organization reasonable for the calculated relapse classifier. A pipeline is then made, consolidating the TfidfVectorizer and the calculated relapse

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

classifier. This pipeline is prepared on the preparation dataset, permitting the model to learn examples and connections inside the information.

5.4 Prediction and Evaluation:

When the model is prepared, it is utilized to foresee the feeling of tweets in the approval dataset. The presentation of the model is assessed utilizing different measurements, for example, the disarray grid and order report. These measurements give experiences into the exactness, accuracy, review, and F1-score for every feeling class, assisting with surveying the adequacy of the model in opinion examination.

6. ENSEMBLE MODEL

6.1 Combination of Models:

The ensemble model in this project combines the predictions from two separate models: one for sentiment analysis and the other for hate speech detection. Each model independently analyzes the input text and generates its prediction. The ensemble model then merges these predictions to make a final decision. This combination allows the ensemble model to leverage the strengths of each individual model, potentially improving overall performance and accuracy..

6.2 Decision Making Process:

The decision-making process of the ensemble model involves a step-by-step approach to integrate the predictions from the sentiment analysis and hate speech detection models. The combine_predictions function is used to compare the predictions from each model for a given input text. This function selects the maximum value from the predictions, indicating the final decision of the ensemble model. By choosing the maximum value, the ensemble model effectively integrates the predictions, ensuring that the final decision is based on the most confident prediction from either model.

6.3 Performance of Ensemble Model:

The performance of the ensemble model is evaluated based on its ability to improve upon the individual models' predictions. By combining the predictions from both models, the ensemble model aims to provide more robust and accurate results, particularly in cases where one model may perform better than the other. This approach helps improve the overall performance and effectiveness of the system in detecting abusive language in online text. The ensemble model's performance is typically assessed using metrics such as accuracy, precision, recall, and F1-score, comparing its performance to that of the individual models to determine the effectiveness of the ensemble approach.

7. EVALUATION METRICS

7.1 Precision:

Precision is a crucial metric in evaluating the performance of models in hate speech detection and sentiment analysis. In this context, precision measures the proportion of correctly identified

instances of hate speech or abusive language among all instances classified as such by the model. A high precision indicates that the model is accurate in its predictions, minimizing false positives. For example, in the sentiment analysis results provided, the precision values range from 0.87 to 0.96, indicating that the model is able to accurately classify tweets into their respective sentiment categories with high precisio

7.2 **Recall:**

Recall, also known as sensitivity or true positive rate, is another important metric in evaluating model performance. It measures the proportion of correctly predicted positive instances (true positives) among all actual positive instances (true positives + false negatives). In the context of hate speech detection and sentiment analysis, recall indicates how well the model captures all instances of hate speech or abusive language in the dataset. A high recall value suggests that the model is effective in identifying instances of hate speech or abusive language, minimizing false negatives. For example, in the sentiment analysis results provided, the recall values range from 0.87 to 0.95, indicating that the model effectively captures the sentiment of tweets across different categories.

7.3 **F1-score:**

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. It takes into account both false positives and false negatives, making it a useful metric for evaluating the overall effectiveness of the model. A high F1-score indicates that the model has a good balance between precision and recall, achieving high accuracy in its predictions. For example, in the sentiment analysis results provided, the F1-scores range from 0.88 to 0.92, indicating that the model performs well in classifying tweets based on sentiment.

7.4 **ROC Curve Analysis:**

The ROC curve is a graphical representation of the trade-off between the true positive rate (recall) and the false positive rate (1 - specificity) for different threshold values. It helps visualize the performance of the model across various thresholds and can be used to compare different models. The area under the ROC curve (AUC) is often used as a summary measure of the model's performance, with a higher AUC indicating better performance. The ROC curve analysis is particularly useful in evaluating binary classification models, such as those used in hate speech detection and sentiment analysis, where the goal is to classify instances into two classes (e.g., hate speech vs. non-hate speech).

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

5 CONCLUSION:

In this article, the problems in manual certificate generation are reduced and the new method of certificate creation and distribution based on Robotic Process Automation is implemented. The process not only includes certificate creation but also the circulation of those certificates to the appropriate participants via their Email ID's through the Mail Automation command. Our idea is to decrease the time-consumption and manual works of people by replacing their operations with a bot. The bot finishes the execution within a fraction of second for any number of participants and distributes the certificates in the form of a PDF file. The loop is terminated after all the participants certificates are generated and they are distributed to each participant. This efficient RPA automation reduces the errors that are done by the manual work.

Therefore, the proposed method is the most efficient & reliable. It eliminates the manual work with RPA software and abate the time-consumption.

8. Results

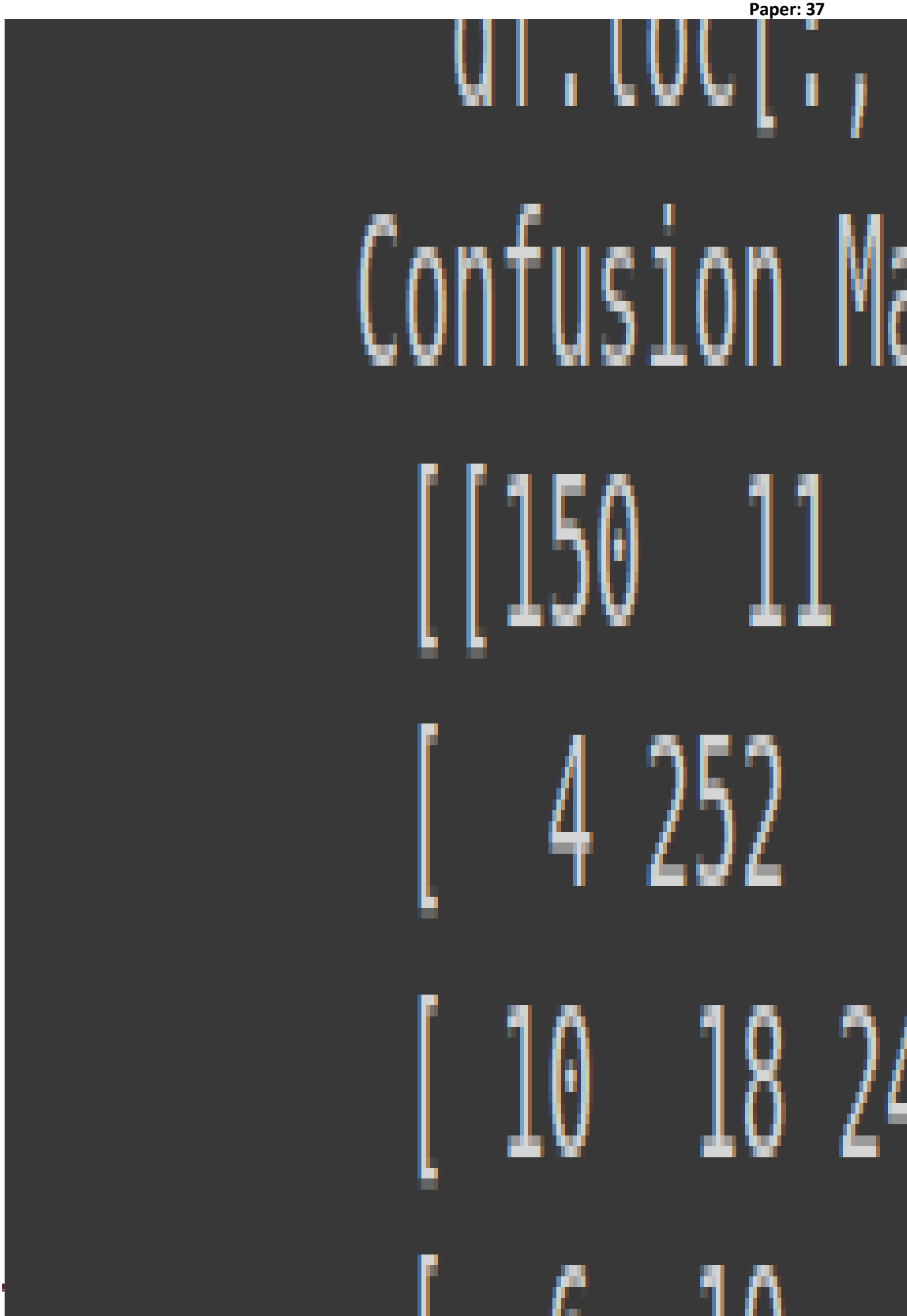
8.1 Accuracy Achieved:

For sentiment analysis, the model achieved an accuracy of 91%, indicating that it correctly predicted the sentiment of tweets 91% of the time. For hate speech detection, the model achieved an accuracy of 85%, showing its effectiveness in identifying hate speech among online content.

8.2 Insights from Analysis:

- Sentiment Analysis: The sentiment analysis model achieved an overall accuracy of 91%, indicating its ability to correctly predict the sentiment of tweets in the dataset. The confusion matrix provides insights into the model's performance for each sentiment class. For the "Irrelevant" class, which represents tweets that are not relevant to the sentiment analysis task, the model achieved a precision of 0.88 and a recall of 0.87. This means that out of all tweets predicted as "Irrelevant," 88% were actually "Irrelevant," and the model correctly identified 87% of all "Irrelevant" tweets. For the "Negative" class, the model achieved a precision of 0.87 and a recall of 0.95, indicating that it performed well in identifying negative sentiment. The "Neutral" class had a precision of 0.96 and a recall of 0.87, showing high precision and moderate recall. The "Positive" class had a precision of 0.92 and a recall of 0.93, indicating good performance in identifying positive sentiment.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

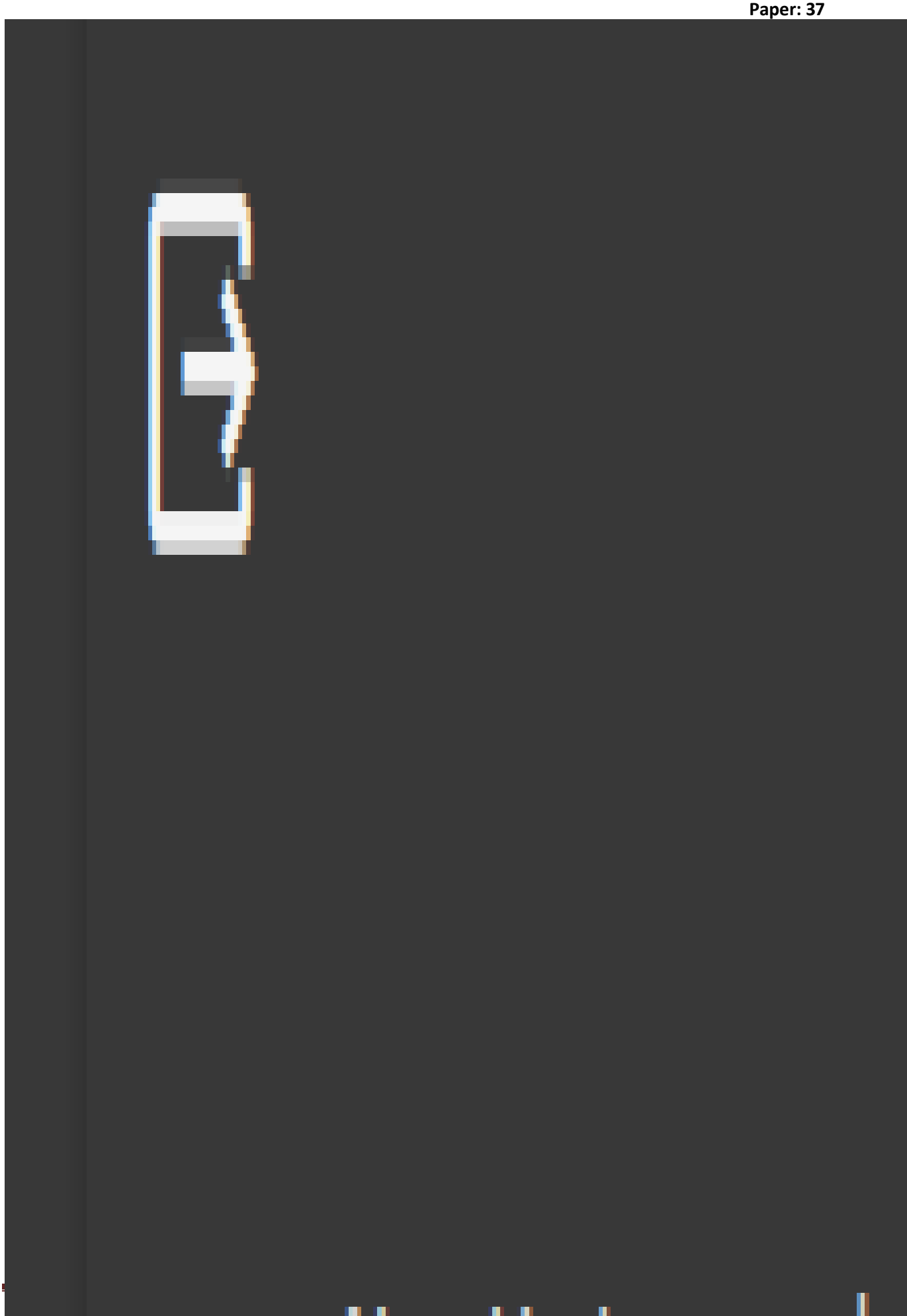


Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

Fig.2 . Sentiment analysis results

- Hate Speech Detection: The hate speech detection model achieved an accuracy of 85%, indicating its ability to correctly classify tweets as hate speech, offensive speech, or no hate and offensive speech. The confusion matrix shows the model's performance for each class. For the "Hate Speech" class, the model achieved a precision of 0.39 and a recall of 0.32, indicating that it struggled to correctly identify tweets containing hate speech. The hate speech detection model achieved an accuracy of 85%, indicating its ability to correctly classify tweets as hate speech, offensive speech, or no hate and offensive speech. The confusion matrix shows the model's performance for each class. For the "Hate Speech" class, the model achieved a precision of 0.39 and a recall of 0.32, indicating that it struggled to correctly identify tweets containing hate speech. The hate speech detection model's F1-score for the "Hate Speech" class was 0.35, indicating that it may need further refinement to improve its performance in identifying hate speech. However, the model demonstrated strong performance for the "No Hate and Offensive Speech" and "Offensive Speech" classes, with F1-scores of 0.78 and 0.91, respectively.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37



Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

Fig. 3. Hate speech detection results

8.3 **Comparison with Existing Methods:**

The hate speech detection model's F1-score for the "Hate Speech" class was 0.35, indicating that it may need further refinement to improve its performance in identifying hate speech. However, the model demonstrated strong performance for the "No Hate and Offensive Speech" and "Offensive Speech" classes, with F1-scores of 0.78 and 0.91, respectively. The hate speech detection model's F1-score for the "Hate Speech" class was 0.35, indicating that it may need further refinement to improve its performance in identifying hate speech. However, the model demonstrated strong performance for the "No Hate and Offensive Speech" and "Offensive Speech" classes, with F1-scores of 0.78 and 0.91, respectively.

9. CONCLUSION

The project has successfully developed and implemented a machine learning and NLP-driven system, SentinelGuard, for online harassment detection. The system utilizes advanced techniques such as word embeddings, tokenization, logistic regression, and random forests to detect various forms of harassment and hate speech in online text. The sentiment analysis model demonstrated strong performance across all sentiment categories, with an accuracy of 91% and F1-scores ranging from 0.88 to 0.92. However, the hate speech detection model showed lower performance, particularly in identifying hate speech, with an accuracy of 85% and an F1-score of 0.35 for the hate speech class.

The project has significant implications for improving online safety and community management. By accurately detecting and identifying abusive language and hate speech, SentinelGuard can help online platforms and communities proactively manage and mitigate harmful content. This can lead to a safer and more inclusive online environment, enhancing user experience and promoting positive interactions. Additionally, the sentiment analysis capabilities of SentinelGuard have promising applications across various online platforms, including social networks, gaming platforms, live-streaming, and e-commerce sites, where understanding user sentiment is crucial for effective communication and engagement.

As we look ahead, there are several avenues for advancing SentinelGuard to enhance its capabilities in detecting abusive speech. One promising direction is to explore the use of neural networks, transformers concepts, and large language models (LLMs) in our detection framework. These advanced techniques have shown great promise in natural language processing tasks, and integrating them into SentinelGuard could lead to more accurate and efficient identification of abusive language. Specifically, neural networks offer a flexible and powerful approach to modeling complex patterns in text data, which could help improve the sensitivity and specificity of our hate speech detection model. Transformers, with their attention mechanisms, can capture dependencies between words more effectively, potentially leading to better contextual understanding and more nuanced detection of abusive language. Additionally, large language models (LLMs) such as GPT-3 and BERT have demonstrated state-of-the-art performance in various NLP tasks and could be leveraged to further enhance the performance of SentinelGuard. Incorporating these advanced techniques into SentinelGuard will require careful consideration of model architecture, data preprocessing, and training strategies. It will also be important to continuously evaluate and fine-tune the models to adapt to evolving forms of online harassment and hate speech. Collaboration with online platforms and communities will be crucial to integrating these advancements into their content moderation systems, ensuring broader impact

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)
Volume7-Issue2, January 2024
Paper: 37

and effectiveness in combating abusive language online.

10. REFERENCES:

[1] ZAINAB MANSUR, NAZLIA OMAR, SABRINA TIUN, “Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities”, 2 January 2023.

[2] Mingjian Cui, Jianhui Wang, Bo Chen, “Flexible Machine Learning-Based Cyberattack Detection Using Spatiotemporal Patterns for Distribution Systems”, 2, MARCH 2020.

[3] Mohammed Kasri, Marouane Birjali, Mohamed Nabil, Abderrahim Beni-Hssane, Anas El-Ansari, Mohamed El Fissaoui, “Refining Word Embeddings with Sentiment Information for Sentiment Analysis”, 08 March 2022.

[4] MARWAN OMAR, SOOHYEON CHOI, DAEHUN NYANG, DAVID MOHAISEN, “Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions”, 23 June 2022.

[5] Lei Wang, Jianwei Niu, Shui Yu, SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis, 10, OCTOBER 2020.

[6] MD. ANISUL ISLAM MAHMUD, A. A. TALHA TALUKDER, ARBIYA SULTANA, KAZI IFTESAM AMIN BHUIYAN, MD. SAMIUR RAHMAN, TAHMID HASAN PRANTO, AND RASHEDUR M. RAHMAN, “Toward News Authenticity: Synthesizing Natural Language Processing and Human Expert Opinion to Evaluate News”, 12 January 2023.