RESEARCH & INNOVATION DIVISION
IJIST

Identification of Fake Contents Using Text-mining Techniques

Saqlain Sajjad[1], Hafiz Muhammad Ghazi[2], Muhammad Asgher Nadeem[3], Muhammad Irfan Habib[4], Muhammad Salman Saeed,[8] Syed Ali Hasnain Naqvi[5], Zeeshan Ahmad Arfeen[6], Isheeaq Naeem[7], Muhammad Irfan[1]

[1]Department of Computer Science, University of Management and Technology Sialkot Campus, Pakistan.

[2]Department of Information Engineering Technology, National Skills University Islamabad, Islamabad, 44310, Pakistan.

[3] Thal University Bhakhar Punjab, Pakistan.

[4]Department of Electrical Engineering Technology, National Skills University Islamabad, Islamabad, 44310, Pakistan.

[5]Faculty of Social Sciences, Sir Syed University of Engineering and Technology (SSUET), Karachi, Pakistan.

[6]Department of Electrical Engineering, The Islamia University of Bahawalpur (IUB), Bahawalpur, 63100 Pakistan.

[7] University of Management and Technology, Sialkot, Pakistan.

[8]Multan Electric Power Company (MEPCO), Multan, Pakistan

*** Correspondence.** zeeshan.arfeen@iub.edu.pk

Citation | Sajjad. S, Ghazi. H. M, Nadeem. M. A, Habib. M. I, Saeed. M. S, Naqvi. S. A. H, Arfeen. Z. A, Naeem. I, Irfan. M, "Identification of Fake Contents Using Text-mining Techniques", IJIST, Vol. 06 Issue. 04 pp 2084-2103, Dec 2024

Received | Nov 11, 2024 **Revised** | Dec 9, 2024 **Accepted** | Dec 14, 2024 **Published** | Dec 19, 2024.

In recent years, social media users have become increasingly concerned about sharing content that may be unpleasant or harmful. The widespread use of platforms like Facebook and Twitter has contributed significantly to this growing awareness. The primary objective of our approach is to accelerate and automate the detection of offensive content posted on these platforms, simplifying the process of taking necessary actions and filtering harmful communications. A benchmark dataset, OLID 2019 (Offensive Language Identification Dataset), is available online to aid in this task. Our study focuses on identifying whether a tweet is offensive. Our team, which included several members, rigorously compared various feature extraction methods and model-building algorithms. Ultimately, our comparative analysis revealed that decision trees were the most effective model. The decision trees applied to the normalized dataset resulted in an 84% improvement in the Macro F1 score, which aligns with previous research. In conclusion, a real-time system could be developed across multiple social media platforms to detect and evaluate objectionable posts, enabling timely interventions to promote healthier online behavior and foster a positive societal impact.

Keywords. Fake Content; Text Mining; Identifications; Text Analysis and Techniques.

Introduction.

In today's world, people communicate and share information through platforms like Facebook, Twitter, and other social networking apps. While sharing content has become easier, it has also led to a rise in fake news, offensive material, and misleading information, creating significant issues for public discourse, media trust, and societal norms. As users become more accustomed to encountering irrelevant or false content, there is a growing need for effective methods to detect and address these problems. This paper applies text-mining techniques to identify and categorize fake information spread on social media, aiming to improve the quality of information on these platforms. By using various datasets and algorithmic approaches on tagged posts, this research seeks to help automate the identification of misleading or offensive content.

Social media platforms allow users to freely express their thoughts and opinions, but to avoid being perceived as rude, it is important to share valuable information. Over the years, the publication of profane or offensive content on social media has increased, raising concerns about the impact of platforms like Facebook and Twitter. More and more people use social media to share details about their lives, and by reading comments, it is easy to gauge whether individuals have positive or negative sentiments on a particular topic. This freedom of expression has both positive and negative effects, enabling people to voice opinions on issues that matter to them and society. However, the rise in popularity of these platforms has also led to challenges such as cyberbullying, harassment, and the spread of hateful speech, which can have severe consequences, including affecting people's mental health or leading to self-harm.

Social media platforms, including Twitter, are frequently studied for their use of abusive language. Various methods have been proposed to detect foul language in social media posts, including neural network-based techniques for analyzing tweet polarity and identifying inappropriate words. For instance, BERT has been used to assess tweet sentiment and detect offensive language, while LSTM models with attention mechanisms have been employed to improve the accuracy of language detection. Additionally, multi-classification techniques have been applied to identify objectionable content, and deep learning has been used to detect offensive language in videos and textual posts across different platforms.

Machine learning approaches, such as feature extraction methods like Bag of Words, N-grams, Chi-square, and TFIDF, have also been utilized to detect inappropriate language in social media posts. Researchers have faced challenges in detecting offensive content due to issues with feature extraction and selection, despite the variety of methods available. As social media platforms like Facebook and Twitter continue to grow, so does the frequency of abusive or harmful language. This has led to concerns about its impact on individual and community mental health. AI, deep learning, and natural language processing techniques, along with feature extraction methods, are being applied to tackle the challenge of identifying harmful language on social media.

In our study, we highlighted the issues related to offensive language identification and the characteristics of the OLID dataset. Since this dataset is relatively new, we set specific goals for our research. Once the baseline results for the dataset are published, we aim to contribute meaningfully to this area of study. Our solution is structured around five key components. text pre-processing, feature extraction, imbalance management, model building, and model evaluation. This framework will help streamline the process of detecting and addressing offensive content, facilitating moderation on social media platforms. OLID 2019, the Offensive Language Identification Dataset, will be used in this research, which focuses on determining whether a tweet contains offensive language.

Novelty of Study. This research is innovative in its use of advanced text-mining techniques, including sentiment analysis and natural language processing, to enhance the detection of fake content on social media. The study provides a comprehensive comparative analysis of various machine learning classifiers, highlighting the effectiveness of the Passive-Aggressive model alongside traditional classifiers like Naive Bayes and Support Vector Machines (SVMs). A key focus of the research is on the ability to accurately identify the sentiment of content. While numerous studies have been conducted on fake news detection, few have specifically addressed sentiment analysis. Moreover, this work tackles the challenge of data availability by advocating for the creation of comprehensive datasets for fake news detection. Additionally, it offers practical insights into social media regulation, proposing automated systems to detect and classify fake news or misleading content. Collectively, these contributions will advance the field of fake news detection and provide a foundation for future research in this area.

Research Question

How can text-mining techniques be effectively applied to identify and classify fake content on social media platforms, and which algorithms and feature extraction methods are most effective for this task?

Objectives of the Study.

The objectives of this study are to.

- Explore and analyze various text-mining techniques for detecting fake content on social media.
- Evaluate the effectiveness of different feature extraction methods in enhancing the accuracy of fake content detection.
- Compare the performance of multiple classification algorithms in the context of identifying fake content.
- Develop a model to automate the identification of offensive and misleading posts using publicly available datasets, such as the OLID 2019 (Offensive Language Identification Dataset).
- Provide recommendations for improving content regulation on social media platforms based on the study's findings.

Relevant Works.

Social media platforms like Facebook and Twitter allow unrestricted communication, but with this freedom has come an increase in offensive and misleading content. As people grow more accustomed to posting such material, concerns have emerged regarding the impact on public discourse and the credibility of media. The proliferation of fake news is particularly troubling, as it has a direct effect on societal behavior, trust in media, and general communication.

In the context of fake news detection, deception has become a major concern for a significant portion of the public. Artificial intelligence (AI) and semantic analysis are being employed to address this issue. For instance, by comparing various machine learning models such as Naive Bayes and random forests, one study found that a random forest classifier achieved a 95.66% accuracy rate in identifying fake content, with bigrams proving more effective than unigrams or trigrams in determining the authenticity of information.

Another approach focuses on classifying news based on its headline, contrasting it with full article classification for accuracy. The goal here is to strike a balance between the speed of information analysis and the precision of fake news classification, using Natural Language Processing (NLP) techniques for both headlines and article content. Such methods often require complex ensemble systems to achieve optimal classification results, though they are effective in improving accuracy.

The difficulty of distinguishing truth from falsehood is further exacerbated by the growing volume of information online. This has led to a surge in research focused on fake news detection, using machine learning classifiers like Naive Bayes, Support Vector Machines (SVM), and semantic analysis. These studies emphasize the importance of language patterns and network analysis for more accurate identification of fake content.

Given the rise of social media, which allows for the rapid spread of information, detecting fake news has become a major challenge. Unlike traditional media, where fake news detection algorithms can be straightforward, the unstructured and noisy data produced by social media complicates the process. Researchers are exploring various data mining techniques to better understand how social media platforms can be used to detect fake news.

Many efforts have focused on improving the precision of fake news detection through machine learning and NLP. One study demonstrated that using AI-driven models for fake news classification could significantly reduce human involvement in the process, increasing both efficiency and accuracy. The use of deep learning and feature-based analysis, including techniques like generative adversarial networks and bidirectional transformers, is being explored to address the growing issue of fake news.

Moreover, the complexity of fake news detection requires continuous refinement of classification systems. As digital content continues to grow, traditional approaches are often rendered ineffective, and new methodologies are needed to keep up with the changing landscape of social media and online content. By advancing machine learning and text analytics, researchers are contributing to the development of more accurate, automated systems for identifying and classifying fake news across platforms.

Table 1. LSTM-Based Previous Studies

Reference	Dataset	Technique	Accuracy	Limitations
[31]	OLID	LSTM	89.5%	No features extraction
[32]	Hate Speech	LSTM	82%	Target of offensive language
[33]	Hate Speech	LSTM	84.5%	Target of offensive language
[34]	Hate Speech	LSTM	85.565%	No Feature Engineering

Table 2. BERT-Based Previous Studies

Reference	Dataset	Technique	Accuracy	Limitations
[35]	Hate Speech	BERT	82%	Target of offensive language
[36]	Hate Speech	BERT	90.4%	No Feature Engineering

Table 3. CNN-Based Previous Studies

Reference	Dataset	Technique	Accuracy	Limitations
[37]	OLID	CNN	80.9%	Target of offensive language
[38]	OLID	CNN	81.2%	Target of offensive language
[39]	OLID	CNN	82.24%	Target of offensive language

Table 4. RNN-based previous studies

Reference	Dataset	Technique	Accuracy	Limitations
[40]	OLID	RNN	82.4%	Target of offensive language
[41][42]	OLID	RNN	81.5%	Target of offensive language

Proposed Methodology.

The OLID dataset is designed to identify offensive language on social media platforms like Twitter and is publicly available for research purposes. It was created by collecting tweets using the Twitter API, with a focus on content that potentially includes offensive language. The dataset categorizes offensive language into three main tasks.

1. **Sub-task A**. Determining whether a tweet contains offensive language (binary classification).
2. **Sub-task B**. Classifying the type of offense (e.g., personal attack, group attack).
3. **Sub-task C**. Identifying the target of the offensive language (e.g., individual, group, or other).

This structure allows for a detailed analysis of offensive language in social media contexts.

Preprocessing Steps.

Before analysis, the dataset underwent several preprocessing steps to ensure the quality and consistency of the data.

- **Tokenization**. Breaking down tweets into individual words or tokens.
- **Lowercasing**. Converting all text to lowercase for uniformity.
- **Removing Special Characters**. Eliminating punctuation and non-essential symbols that do not contribute to the meaning.
- **Stopword Removal**. Filtering out common, unimportant words (e.g., "and," "the") that do not add significant value to the analysis.
- **Lemmatization**. Reducing words to their base or root form to standardize variations.

The OLID dataset further categorized items by source to help identify and classify potentially toxic social media posts. The data was sourced from Twitter, with 14,100 original tweets. After preprocessing, 13,240 tweets were used for training and testing. Tweets were categorized into one of three target groups (individual, group, or other) based on their offensive/non-offensive nature and the identified target audience. The sequence of analysis phases reflected these relationships. Whether a tweet is offensive or not, it could still have a target or none at all. If the tweet is offensive, it could target an individual, a group, or any other entity. This dataset was used in the SemEval-2019 and OffensE-val-2019 challenges.

Dataset Descriptions.

The OLID dataset uses a hierarchical annotation system, where up to three labels can be assigned to a single instance, each representing one of the following levels.

- **Sub-task A**. Identifying offensive language.
- **Sub-task B**. Automated classification of offenses.
- **Sub-task C**. Identifying the target of the offense.
- **NULL**. Assigned when a label is not specified (e.g., INSTANCE NOT NULL NULL).

The dataset file contains the following columns.

- **ID**. Unique identifier for each tweet.
- **Tweets**. The actual tweet content.
- **Subtasks A, B, and C**. Representing the annotation for offensive language identification, offense classification, and offense target identification, respectively.

The following is a list of possible labels for the annotation.

Labels and Tasks.

Level A. Identifying offensive words.

- This communication does not contain any profanity or offensive language.
- The post includes language that could potentially offend some readers, either subtly or overtly.
- If a post contains profanity or any specific offense, whether implied or explicitly stated, it is labeled as "offensive" (OFF) in the annotation.

Level B. Classification of Offences Automatically at Level B

"Targeted insults and threats" refer to insults or threats aimed at a specific individual, group, or organization, as defined in sub-task C categories. An insult may or may not be directed at a particular person or group, depending on the context. While profanity is not the primary focus, it is still considered inappropriate.

Level C. Identifying the offensive target.

An offensive post is categorized as directed at an "individual" if it targets a specific person, such as a celebrity, someone mentioned in the discussion, or an anonymous participant. "Groups" refer to collective entities such as ethnic, gender, or sexual orientation groups, political parties, religious denominations, or any other group sharing common values.

The label "OTH" is used when the targeted audience does not fit into the aforementioned categories, such as an organization, situation, event, or issue. Combination labels may also be used in cases where multiple categories apply.

Table 5. Dataset Distribution for Each Task

Sub-task	Total Tweets	Offensive	Non-offensive	Personal Attack	Group Attack	Other
A	13,240	6,000	7,240	2,500	2,500	1,700
B	13,240	6,000	-	2,500	2,500	1,700
C	13,240	6,000	-	2,500	2,500	1,700

The OLID annotation supports various label combinations. The label "Tin" (ind|grp|oth) is considered valid as long as it is not null. It is only marked as null when the tin itself is null.

Dataset Attribute Statistics.

The following figures present the statistics for each task and its corresponding sub-tasks.

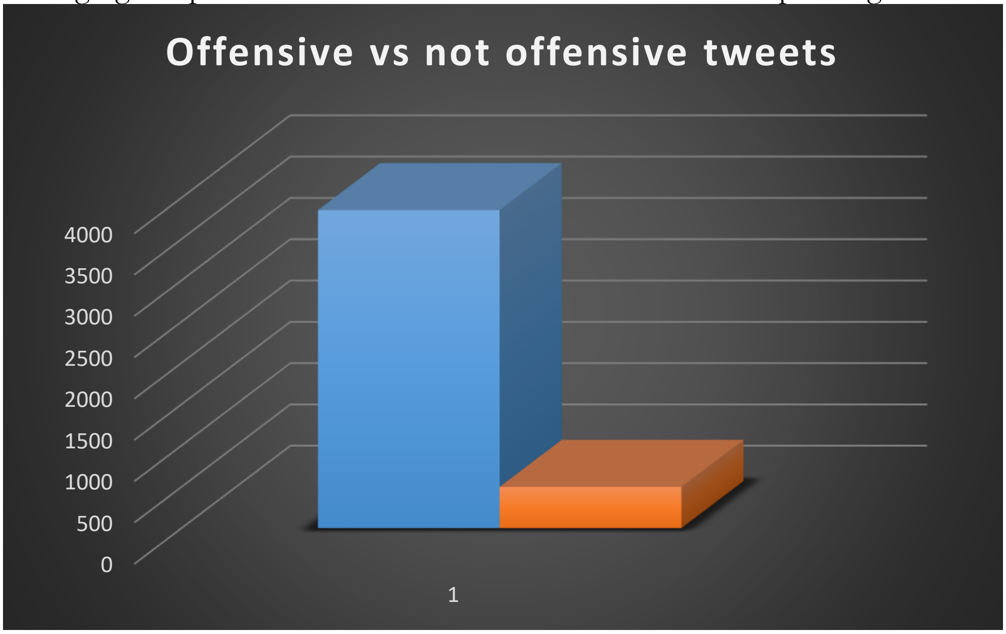

Figure 1. Offensive vs not offensive tweets.

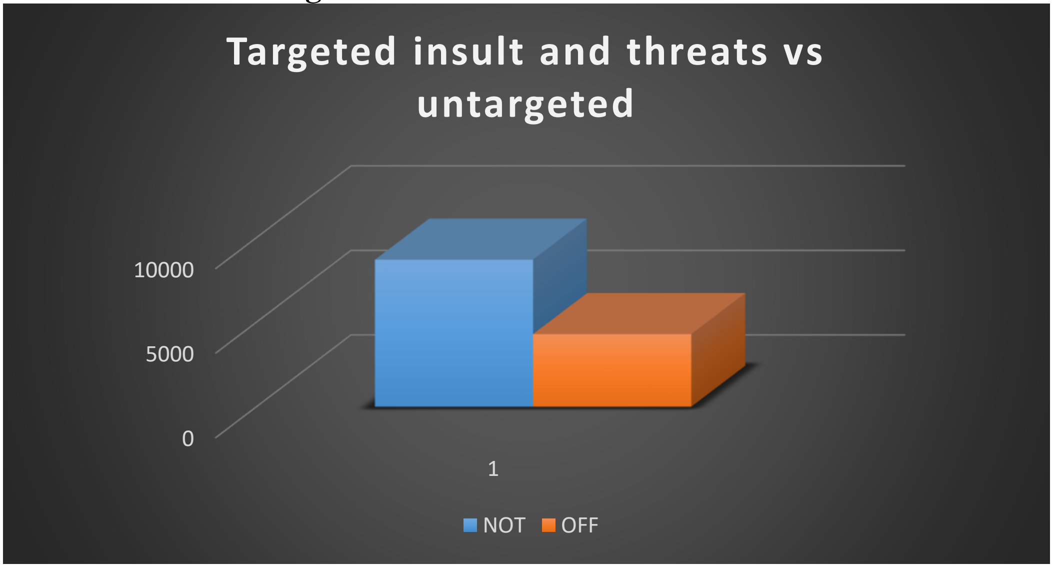

Figure 2. Targeted insult and threats vs untargeted.

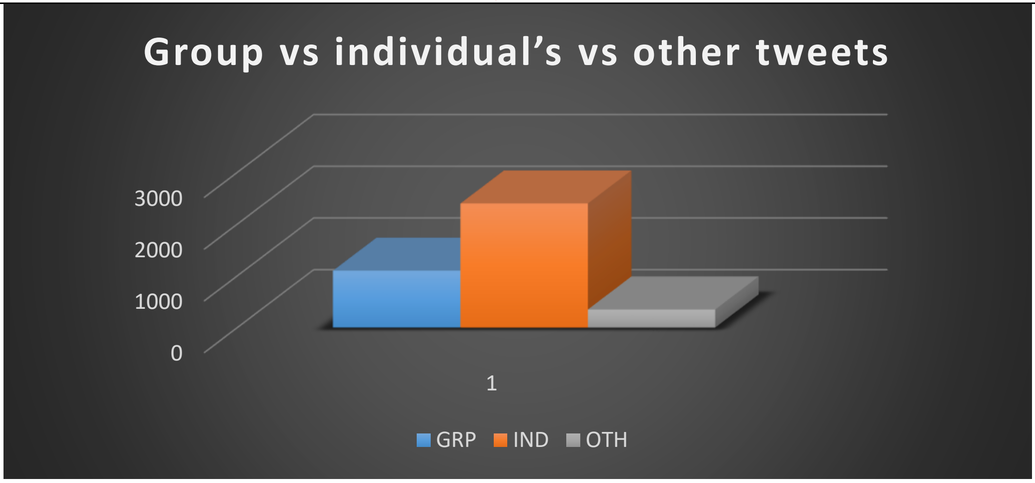

Figure 3. Group vs individual vs other tweets.

Text Data Processing.

The data underwent cleaning after text pre-processing to ensure its suitability for model use. The raw text contained noise such as sentiment, punctuation, varied cases, and other irrelevant elements. A significant part of our NLP work involved preparing the text for easier processing by computer algorithms. Tokenization was one of the methods used to prepare the text, along with a window function applied at the base. Unnecessary words and phrases were removed, making the text more manageable for analysis. Overall, pre-processing played a central role in refining the data, ensuring that only relevant information remained.

Feature Engineering.

In terms of feature engineering, the study extracted several key features, including.

- **N-grams**. Sequences of words that capture context and common phrases associated with offensive language.
- **Sentiment Scores**. Analysis of the emotional tone of tweets to detect negative sentiments often linked to offensive content.
- **Word Frequency**. Measurement of specific words or phrases that indicate offensive language.

These features contribute to enhancing the performance of the classification tasks by providing linguistic cues and sentiment information, which improve the accuracy of offensive language detection. Through this process, domain-specific data can be leveraged to build AI capabilities. Data science heavily relies on the extraction and interpretation of data into AI representations, such as understanding how specific variables are interconnected through a correlation framework.

Correlation Matrix.

A covariance network is similar to a correlation network. By calculating the correlation, we can identify the strength of a linear relationship between two variables. Correlation is a statistical term that indicates how frequently and in what direction two numerical factors are related by a straight line.

Cross-Validation.

A small sample of data is used to refine models in machine learning through a process known as cross-validation. This technique involves using a single variable to determine how many data subsets can be generated from the sample. It is commonly referred to as "K-fold cross-validation."

K-Folds Cross Validation. By using a value between 5 and 10, the dataset is randomly divided into K folds, depending on the amount of data available. The model is then validated using K-1 folds, with the final Kth fold used for testing.

Classification Algorithms. The OLID dataset utilizes five effective classifiers for predicting offensive language. SVM, KNN, Decision Trees, Random Forests, and Logistic Regression.

Support Vector Machine (SVM). Support Vector Machine (SVM) uses a hyperplane to classify data by first assessing the problem's dimensionality. To balance data dimensions, SVM can reduce the number of dimensions in a dataset. The margin distance, which is the gap between classes, is calculated to optimize the separation. Key SVM parameters, such as the kernel, C coefficients, and intercepts, are chosen to suit the data. Kernels, particularly linear and Gaussian (RBF), are essential in tuning SVM for specific datasets, with linear kernels being ideal for linearly separable data. For text classification, data must be converted into a vector format. A major advantage of SVM is its ability to automatically determine optimal parameters, reducing the need for manual tuning. Due to its simplicity and high success rate, SVM, especially the Linear Support Vector Machine, is considered one of the most effective methods for text classification.

K-Nearest Neighbors (KNN). The K-Nearest Neighbors (KNN) algorithm is used to solve both regression and classification problems. It works by comparing new data points to existing ones and discovering patterns based on distance functions. A common method for classifying neighbors is through a majority vote. KNN identifies objects by their features in relation to nearby training samples, predicting a class based on the labels of the nearest matches. Traditionally, researchers used Euclidean distance to determine the nearest neighbor. According to the KNN algorithm, new cases are classified based on their proximity to previously recorded cases. Since the 1970s, KNN has been a widely used non-parametric method for statistical estimation and pattern recognition. Despite its simplicity, KNN often produces impressive results and can also be applied to solve regression problems.

Logistic Regression (LR). Regression analysis is used to categorize data into statistically significant groups. In logistic regression, the outcome variable is typically binary, such as yes or no. For example, 1 indicates the presence of diabetes, and 0 indicates its absence. Unlike traditional regression models, logistic regression is a linear model and is therefore referred to as a log-linear classifier. This model can predict the outcome of a single test, allowing for the observation of its logistic value. If the data is not normally distributed, it is assumed to follow a Gaussian distribution. Logistic regression has multiple teaching methods for handling various attributes and has often been underutilized in some areas. Despite this, it is particularly effective for identifying conditions such as the presence of a disease. Logistic regression has historically been a widely used machine learning algorithm.

Although simple in design, it is powerful in its application. It is especially useful for analyzing binary variables and examining the relationships between binary data. In this study, logistic regression is applied for binary classification, using a 70/30 split of the dataset for training and testing. The logistic regression classifier computes a weighted sum of input features, which is then passed through a sigmoid function. This function converts a real number into a binary outcome (0 or 1), thereby estimating the relationship between variables. Due to its simplicity and effectiveness, logistic regression remains a popular choice for many binary classification tasks.

Random Forest (RF).

Random Forests, an ensemble learning method, constructs multiple decision trees for classification and other tasks. During training, each tree is built independently, and the final classification is determined by averaging the predictions from individual trees. In the context of diabetes classification, the majority vote method is employed, where the condition of a patient—whether healthy or sick—is determined based on the consensus of the decision trees. This approach leverages the combined power of multiple models to improve accuracy and robustness.

Decision Trees.

Decision trees are widely used for both classification and regression tasks, making them essential learning tools in computing. In classification, the decision tree partitions the feature space into distinct regions to categorize input data. For each segment of the dataset, an incremental decision tree is created, which is then split further into smaller segments. The result is a tree-like structure with decision nodes and leaf nodes. Decision trees are particularly effective for handling non-linear data and are commonly applied across various fields, including engineering, law, business, and civil studies. There are two main types of decision trees. categorical and continuous, each suited to different types of data. Additionally, decision trees are useful in processes like backward propagation in machine learning.

Performance Parameters.

The performance of the proposed method is evaluated using metrics such as precision, sensitivity, specificity, and the receiver operating characteristic (ROC) curve. The efficiency of this approach is illustrated in the following graph.

$$\text{Specificity}=\text{TN}/((\text{TN}+\text{FP})) \quad (1)$$

$$\text{Accuracy} =((\text{TP}+\text{TN}))/((\text{TP}+\text{TN}+\text{FP}+\text{FN})) \quad (2)$$

$$\text{Sensitivity} =\text{TP}/((\text{TP}+\text{FN})) \quad (3)$$

$$\text{ROC} = (\text{Sensitivity}+\text{Specificity})/2 \quad (4)$$

Accuracy.

Confirming the accuracy of a model is crucial to ensure its validity. By evaluating the training process, one can quickly determine whether the model is properly trained. A comprehensive assessment was conducted to evaluate the accuracy and effectiveness of each method.

Sensitivity and Specificity.

The ability of a component to correctly identify a positive instance (true positive rate) indicates its responsiveness. Specificity refers to the model's ability to accurately identify individuals who do not have the condition.

Result.

To reduce the dataset size and improve the re-modeling process, we employ text preprocessing techniques. This helps accelerate the training of our model. The OLID dataset, used for offline learning in the prediction task, includes two categories. offensive (OFF) and non-offensive (NOT). Preprocessing will be applied to the tweet section for both training and testing phases. Cleaned data yields more reliable results, as the model can more easily interpret information that contributes meaningfully to its semantics and syntactic structure, as opposed to irrelevant or noisy data. Exploratory data analysis revealed an imbalance in our training set, with 30% of tweets being offensive and 70% non-offensive. To address this, we have explored various over-sampling and under-sampling techniques. If the results favor the majority class, it indicates a significant imbalance that needs correction. Algorithmic demonstration involves developing different AI models to predict new or test data.

Machine Learning Models.

To prepare the data for machine learning, a CSV file was generated. The KNN classifier achieved an accuracy of 73% for Task A, 63% for Task B, and 31% for Task C. In comparison, the SVM classifier achieved 83.8% accuracy for Task A, 75.75% for Task B, and 52.5% for Task C. The Random Forest (RF) classifier achieved 82.8% accuracy for Task A and 66.7% for Task B. The Decision Tree (DT) classifier showed accuracies of 75.4% for Task A, 84% for Task B, and 66% for Task C. These results indicate varying performance across different tasks, with each classifier demonstrating its strengths in specific areas. Various classification methods were employed to detect objectionable language within the text.

KNN.

The default performance of the KNN algorithm is displayed in the figure below.

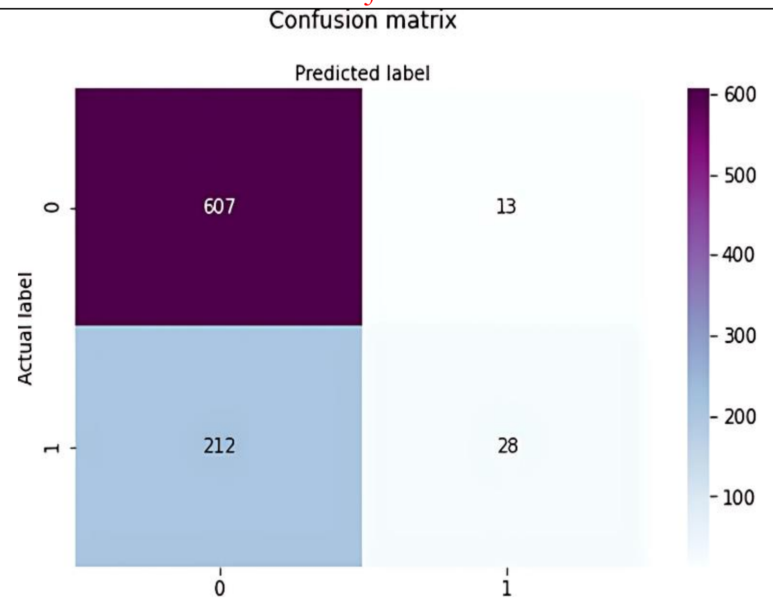

Figure 4. Confusion matrix for Task A.

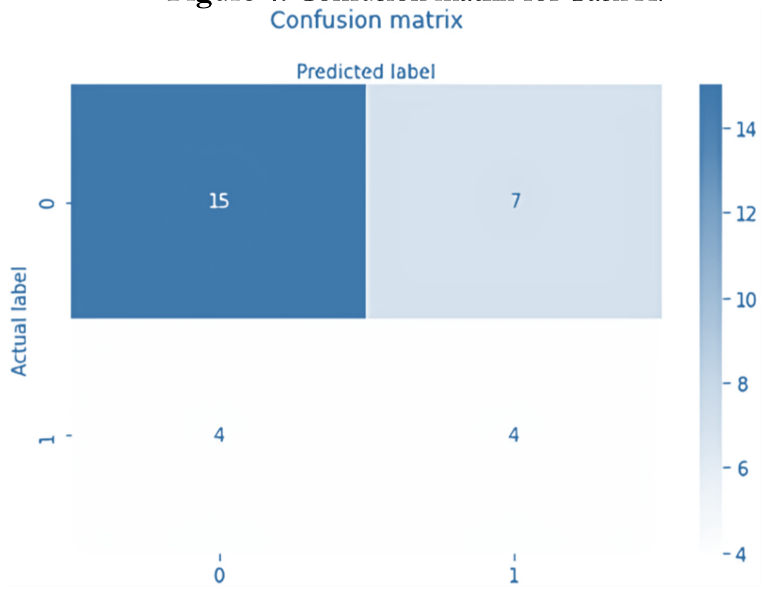

Figure 5. Confusion matrix for Task B.

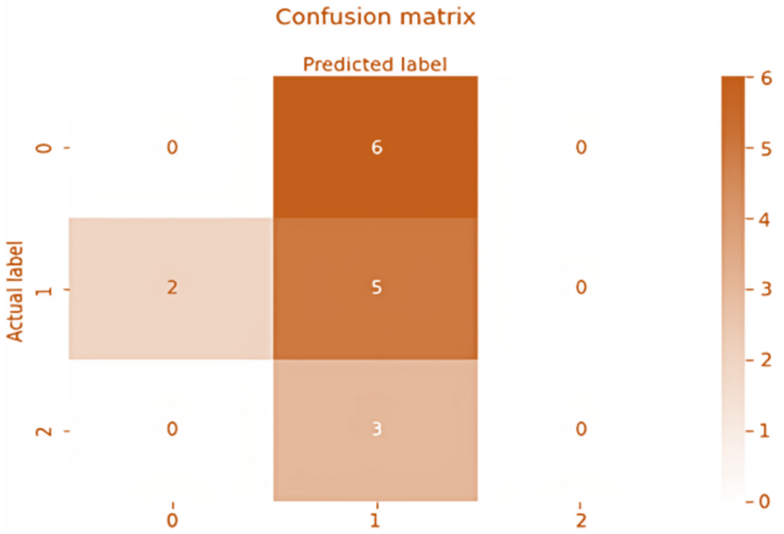

Figure 6. Confusion matrix for Task C.

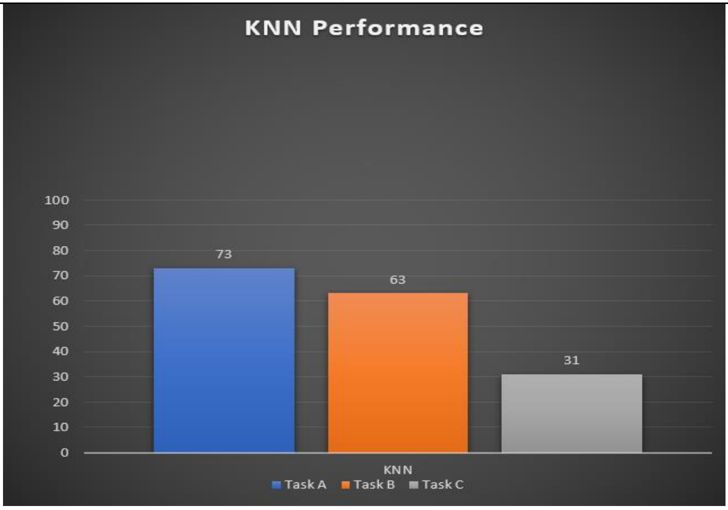

Figure 7. K-Nearest Neighbor's performance.

Figure 7 shows the performance of the KNN algorithm, with 73% accuracy for Task A, 63% accuracy for Task B, and 31% accuracy for Task C.

Logistic Regression.

The following figures illustrate the performance of Logistic Regression.

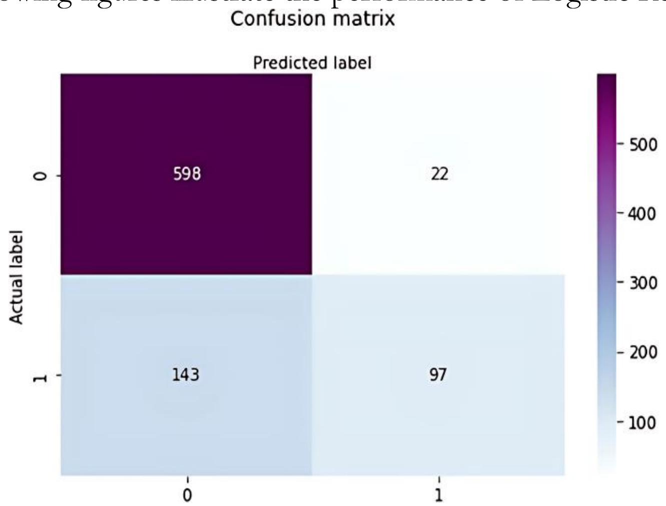

Figure 8. Confusion matrix for Task A.

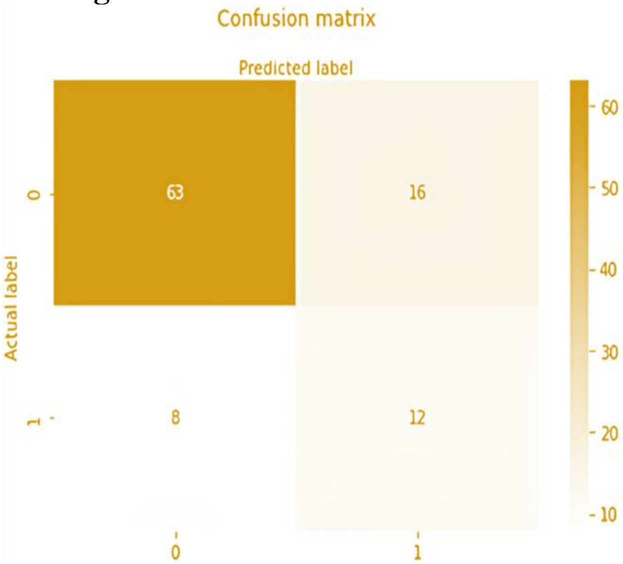

Figure 9. Confusion matrix for Task B.

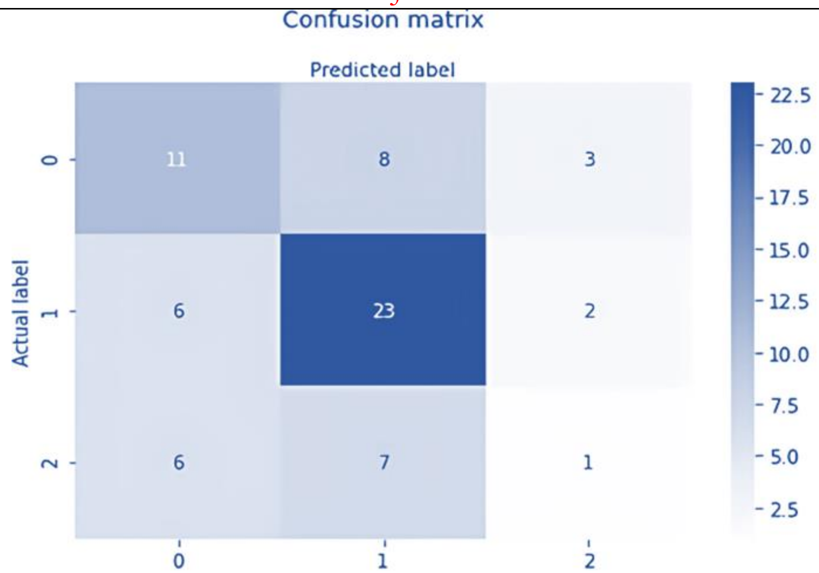

Figure 10. Confusion matrix for Task C.

The following figures present the default performance of the SVM algorithm.

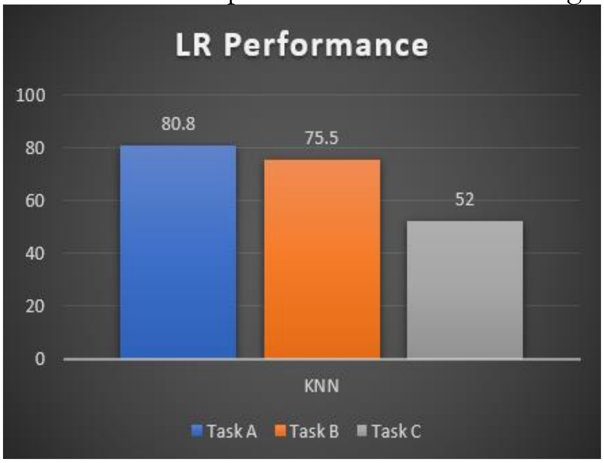

Figure 11. LR Performance.

Figure 11 displays the performance of LR across the tasks. LR achieved an accuracy of 80.8% for Task A, 75.75% for Task B, and 52% for Task C.

Support Vector Machine.

The figures below illustrate the default performance of the SVM algorithm.

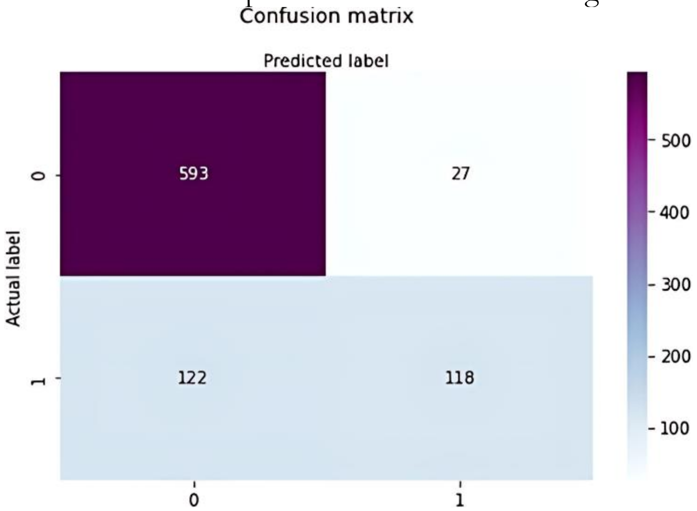

Figure 12. Confusion matrix of Task A.

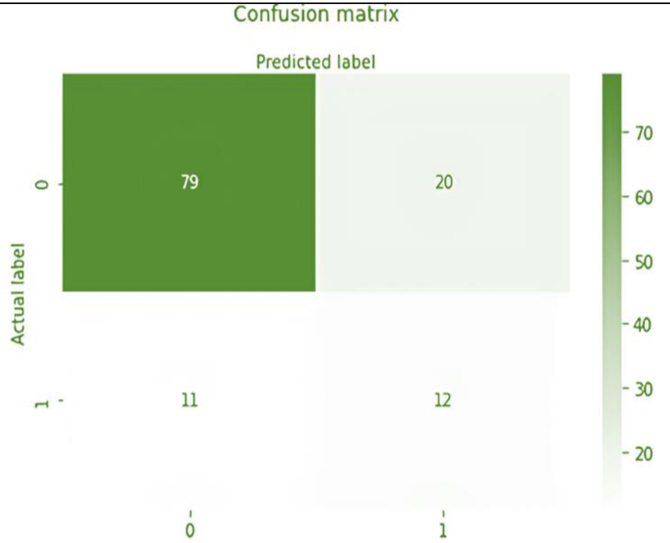

Figure 13. Confusion matrix of Task B.

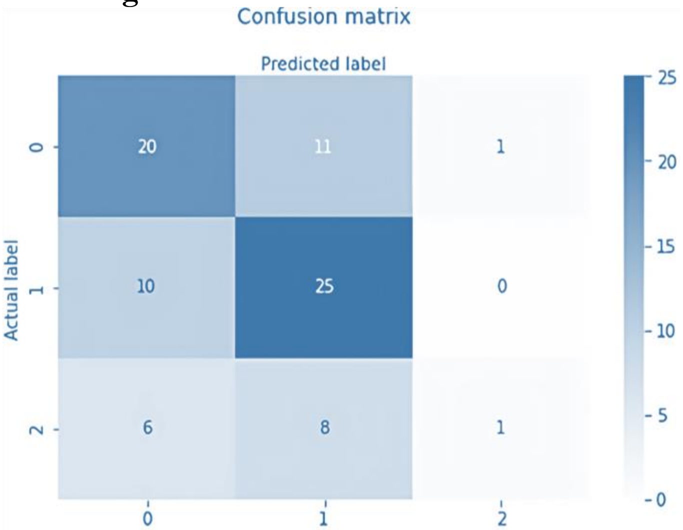

Figure 14. Confusion matrix of Task C.

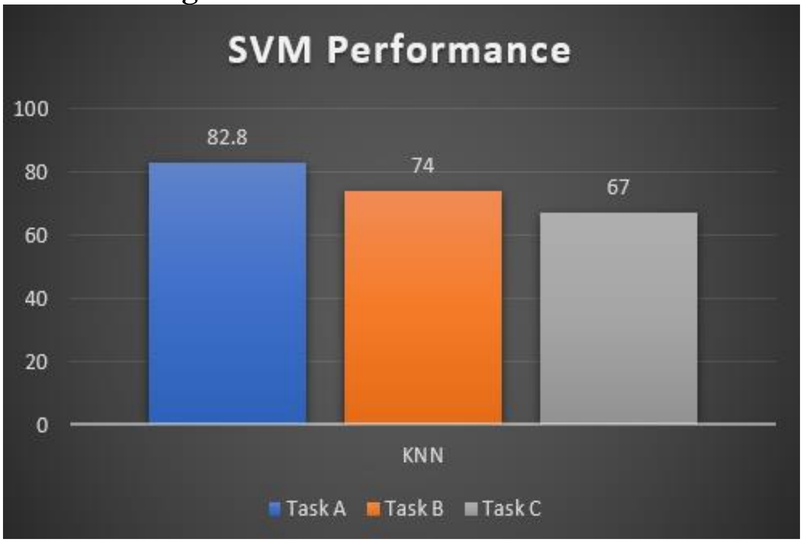

Figure 15. Support Vector Machine Performance.

Figure 15 illustrates the performance of SVM across the tasks. SVM achieved an accuracy of 82.8% for Task A, 74.75% for Task B, and 67% for Task C.

Random Forests.

The RF parameters were set to their default values, as shown in Figure.

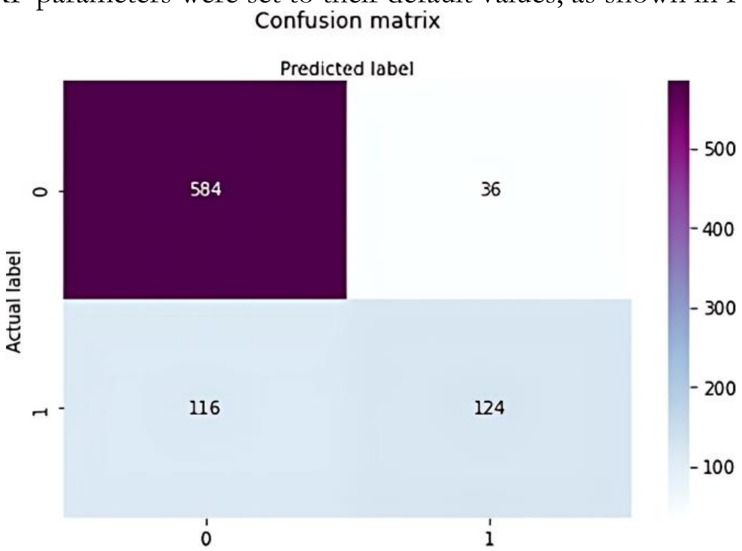

Figure 16. Confusion matrix of Task A.

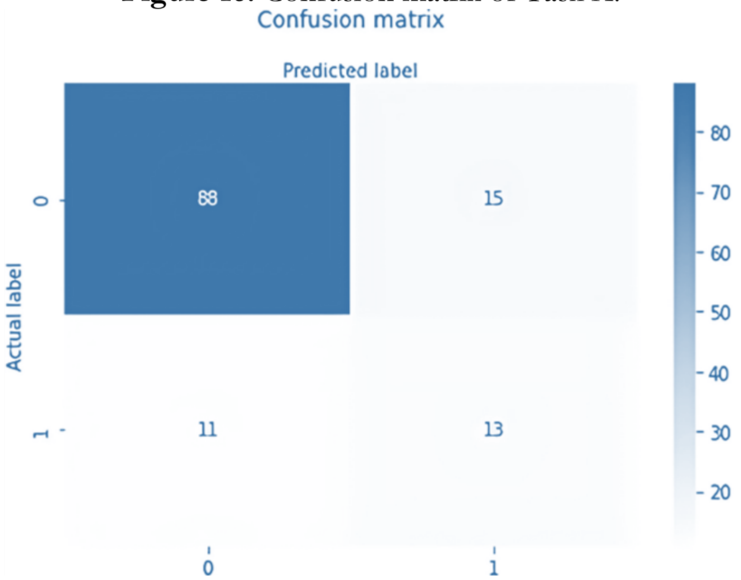

Figure 17. Confusion matrix Task B.

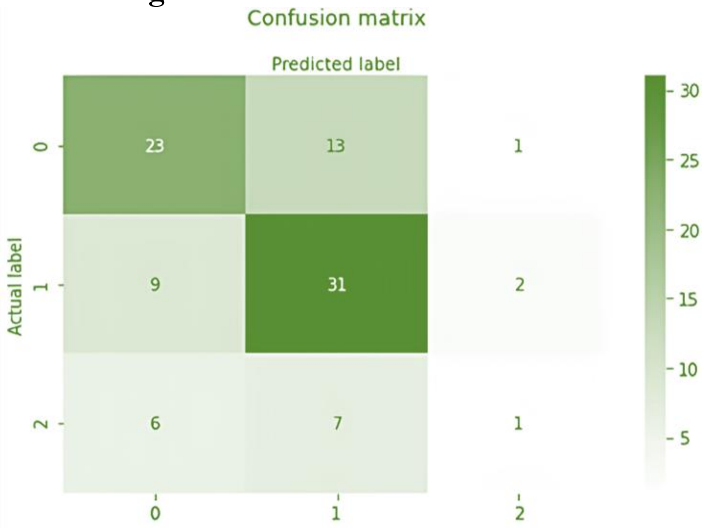

Figure 18. Confusion matrix of Task C.

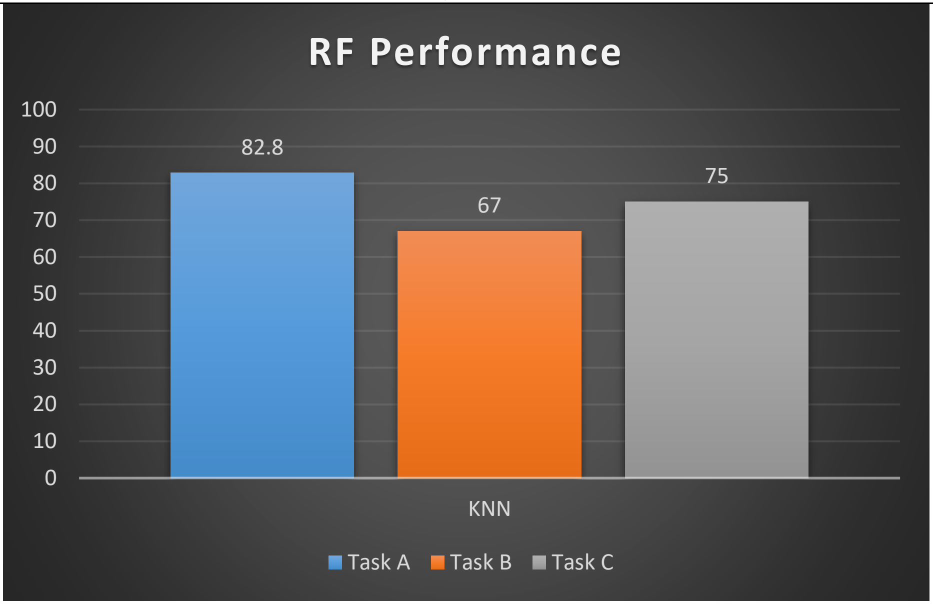

Figure 19. RF Performance.

Figure 19 shows the performance of RF, with an accuracy of 82.8% for Task A, 67% for Task B, and 75% for Task C.

Decision Trees.

As shown in the figure, the Decision Trees algorithm performs well with its default settings.

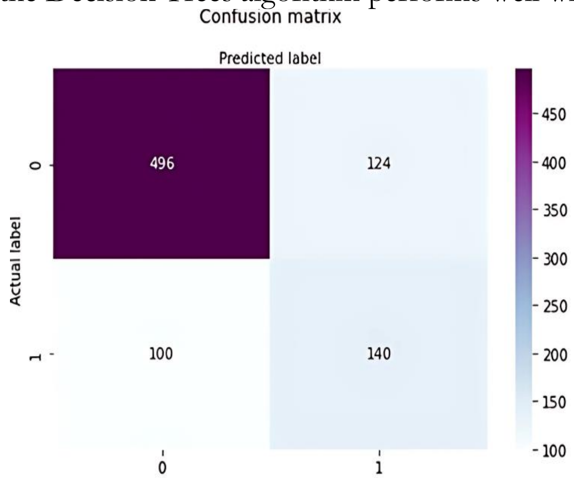

Figure 20. Confusion matrix Task A.

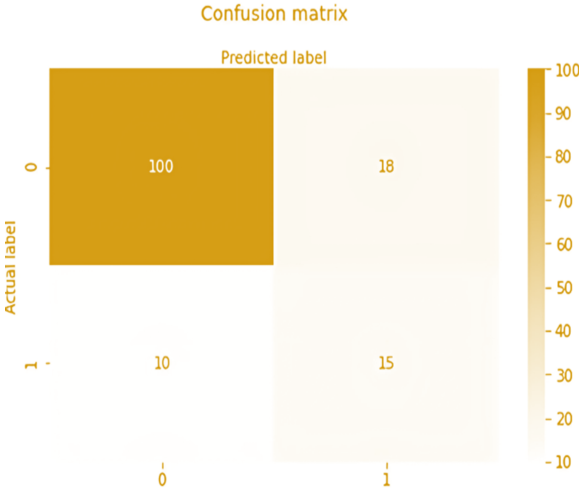

Figure 21. Confusion matrix Task B.

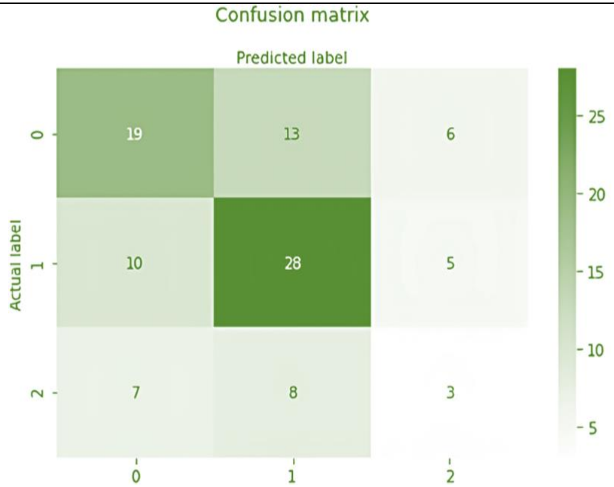

Figure 22. Confusion matrix Task C.

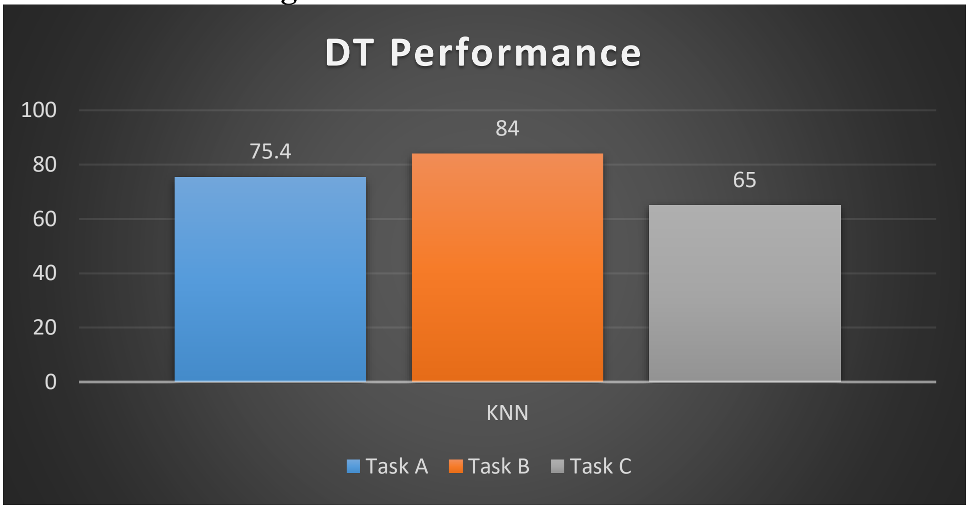

Figure 23. Decision Tree performance.

Figure 23 shows the performance of the Decision Trees (DT) algorithm. it achieved 75.4% accuracy for Task A, 84% for Task B, and 65% for Task C.

Discussion.

This study demonstrated how various text mining techniques and machine learning algorithms can identify fake news and misleading content on social media platforms. The findings indicate that detecting such misinformation is a challenging task, requiring the use of advanced methods to achieve higher accuracy.

Effectiveness of Text-Mining Techniques.

The analysis confirmed that text-mining techniques, particularly those involving word or text combinations related to sentiment and Natural Language Processing (NLP), are effective in distinguishing real from fake content. The results showed that models can achieve higher accuracy when utilizing feature extraction methods that incorporate contextual information. Consistent with previous studies, context-aware models perform better in understanding the language used on social media, making them competitive with global models.

Comparison of Classification Algorithms.

The results of the comparative study on the Passive-Aggressive, Naive Bayes, and Support Vector Machine classifiers were intriguing. While all classifiers showed effectiveness, the Passive-Aggressive classifier outperformed the others in terms of precision and recall. This suggests that the fast-paced nature of social media is particularly well-suited to adaptive learning algorithms that evolve with the data. These findings align with previous studies, such as [20], which have advocated for the use of adaptive models in detecting fake news.

Feature Extraction Methods.

The study found that combining semantic features with other feature extraction methods significantly improves the classification performance of tertiary data sets. Incorporating sentiment analysis as a feature was particularly effective, as it helps the model identify the emotional tone of the content, which often signals its authenticity. The author emphasizes that understanding the intention behind a message is crucial for fake news detection, as highlighted in previous studies [7][8][9][10][11][12][13][14].

Challenges in Data Availability.

One of the key challenges identified in the study was the limited availability of publicly accessible datasets for fake news detection. Relying on a small number of datasets can hinder the generalizability of the models created. This issue highlights the need for a diverse range of data that reflects the broad variety of content found on social media. To advance this field, future research should focus on creating and sharing such datasets, as suggested in earlier studies [18][19][43][20][21][22][23].

Implications for Social Media Regulation.

The findings of this study have significant implications for regulating content on social media platforms. Social media companies should implement automated systems to accurately identify and control the spread of fake content, helping to curb misinformation. The growing body of literature underscores the increasing demand for systemic accountability in the management of online content [1][2][3][4][5][6][7][8][9][10][11][21].

Comparative Discussion.

The OLID dataset is a valuable resource for studying offensive language, offering a diverse range of examples and well-organized categorization. Its key strengths are its comprehensiveness and precise labeling, which enable detailed analysis of each sub-task, such as determining whether a tweet is offensive and identifying its type. However, one potential drawback of social media-based research is its inherent biases. As the dataset is sourced from social media platforms, it may not fully represent all societal groups, limiting its generalizability.

Conclusion.

In recent years, the number of individuals posting offensive or abusive content on social media has surged significantly. The widespread use of platforms like Facebook and Twitter has given rise to numerous challenges due to their immense popularity. The primary goal of this study is to accelerate and automate the detection of offensive content, making it easier to take appropriate actions and regulate harmful messages. The OLID 2019 (Offensive Language Identification Dataset), being publicly available, serves as a starting point for this research. The aim is to determine whether a tweet contains objectionable content based on its text.

To balance the training dataset, we applied the Random Under-sampling technique. Additionally, a comparative study was conducted on various feature extraction methods and model building algorithms. The results showed that KNN achieved 73% accuracy for Task A, 63% for Task B, and 31% for Task C. Logistic Regression (LR) performed with 80.8% accuracy for Task A, 75.7% for Task B, and 52% for Task C. Support Vector Machine (SVM) achieved 82.8% accuracy for Task A, 74.75% for Task B, and 67% for Task C. Random Forest (RF) scored 82.8% for Task A, 67% for Task B, and 75% for Task C. Decision Trees (DT) had 75.4% accuracy for Task A, 84% for Task B, and 66% for Task C.

The final comparative analysis revealed that decision trees performed the best, achieving the highest Macro F1 score of 84%, surpassing previous studies. Our results outperformed earlier works, highlighting the potential for real-time detection and evaluation of offensive content on social media platforms. This could help regulate behavior online and contribute to a more positive digital environment.

Acknowledgment.

The authors would like to thank all contributing institutions.

Author's Contribution.

Conceptualization, HMG, MAN, and Z.A.A.; Methodology- MAN, HMG; Software-HMG, IN; Validation-HMG; Formal analysis, AIHN, MI; Investigation- MAN, IN, MI.; Resources –IN, ZAA, SS; data curation- SIHN.; writing—original draft preparation – SS, IN, MI; writing—review and editing, SIHN.MSS; supervision-SS, ZAA, HMG; Project Administration –SS, ZAA. All authors have read and agreed to the published version of the manuscript.

Conflict of interest.

The authors declare no conflicts of interest.

References.

[1] G. M. S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, "Automatic Hate Speech Detection using Machine Learning. A Comparative Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, 2020, doi. doi. 10.14569/ijacsa.2020.0110861.

[2] H. Ahmed, "Detecting opinion spam and fake news using n-gram analysis and semantic similarity," 2017.

[3] F. C. S. Ahmed, K. Hinkelmann, "(PDF) Development of Fake News Model using Machine Learning through Natural Language Processing," arXiv (Cornell University). Accessed. Dec. 22, 2024. [Online]. Available. https.//www.researchgate.net/publication/357952759_Development_of_Fake_News_Model_using_Machine_Learning_through_Natural_Language_Processing

[4] and A. I. M. ibn S. S. S. Alanazi, M. B. Khan, "Arabic Fake News Detection In Social Media Using Readers' Comments. Text Mining Techniques In Action," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 20, no. 9, 2020, doi. 10.22937/IJCSNS.2020.20.09.4.

[5] and R. D. W. Aldjanabi, A. Dahou, M. a. A. Al-Qaness, M. A. Elaziz, A. M. Helmi, "Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model," *Informatics*, vol. 8, no. 4, p. 69, 2021, doi. https.//doi.org/10.3390/informatics8040069.

[6] and A. I. E. M. M. Khalil, H. M. Ghazi, M. I. Habib, F. Shahzad, "Guideline for Selecting the Right Content Management System (RCMS) for Web Development. A Comprehensive Approach," *J. Comput. Biomed. Informatics*, 2024, [Online]. Available. https.//www.researchgate.net/publication/380151503_Guideline_for_Selecting_the_Right_Content_Management_System_RCMS_for_Web_Development_A_Comprehensive_Approach

[7] and J. V. M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, "Sentiment Analysis for Fake News Detection," *Electronics*, vol. 10, no. 11, p. 1348, 2021, doi. https.//doi.org/10.3390/electronics10111348.

[8] R. A. A. and M. I. E.-K. Ghembaza, "Anti-Islamic Arabic Text Categorization using Text Mining and Sentiment Analysis Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, 2021, doi. DOI. 10.14569/IJACSA.2021.0120889.

[9] and D. H. B. Collins, D. T. Hoang, N. T. Nguyen, "Trends in combating fake news on social media – a survey," *J. Inf. Telecommun.*, pp. 1–20, 2020, doi. https.//doi.org/10.1080/24751839.2020.1847379.

[10] A. and R. Katarya, "Analysis of Online Toxicity Detection Using Machine Learning Approaches," *arXiv (Cornell Univ.*, 2021, doi. 10.48550/arXiv.2108.01062.

[11] and A. G. N. Ashraf, A. Zubiaga, "Abusive language detection in youtube comments leveraging replies as conversational context," *PeerJ Comput. Sci.*, vol. 7, p. 742, 2021, [Online]. Available. https.//peerj.com/articles/cs-742/

[12] J. A. Waqas Haider Bangyal, Rukhma Qasim, Najeeb ur Rehman, Zeeshan Ahmad, Hafsa Dar, Laiqa Rukhsar, Zahra Aman, "Detection of Fake News Text Classification on

COVID-19 Using Deep Learning Approaches," *Comput. Math. Methods Med.*, 2021, doi. https.//doi.org/10.1155/2021/5514220.

[13] P. Bharadwaj and Z. Shao, "Fake News Detection with Semantic Features and Text Mining," *Int. J. Nat. Lang. Comput.*, vol. 8, no. 3, pp. 17–22, 2019, doi. 10.5121/ijnlc.2019.8302.

[14] S. Aphiwongsophon and P. Chongstitvatana, "Detecting fake news with machine learning method," *ECTI-CON 2018 - 15th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol.*, pp. 528–531, Jul. 2018, doi. 10.1109/ECTICON.2018.8620051.

[15] and T. R. M. C. Buzea, S. Trausan-Matu, "Automatic Fake News Detection for Romanian Online News," *Information*, vol. 13, no. 3, p. 151, 2022, doi. https.//doi.org/10.3390/info13030151.

[16] R. Chatterjee, "Profanity detection in social media text using a hybrid approach of NLP and machine learning," 2021.

[17] G. A. De Souza and M. Da Costa-Abreu, "Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata," *Proc. Int. Jt. Conf. Neural Networks*, Jul. 2020, doi. 10.1109/IJCNN48605.2020.9207652.

[18] and D. P. L. A. S. D. Santos, L. F. R. Camargo, "Evaluation of classification techniques for identifying fake reviews about products and services on the internet," *Gestão & Produção*, vol. 27, no. 4, 2020, doi. https.//doi.org/10.1590/0104-530X4672-20.

[19] E. Elmurngi and A. Gherbi, "An empirical study on detecting fake reviews using machine learning techniques," *7th Int. Conf. Innov. Comput. Technol. INTECH 2017*, pp. 107–114, Nov. 2017, doi. 10.1109/INTECH.2017.8102442.

[20] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting Hate Speech and Offensive Language on Twitter using Machine Learning. An N-gram and TFIDF based Approach," Sep. 2018, Accessed. Dec. 22, 2024. [Online]. Available. http.//arxiv.org/abs/1809.08651

[21] E. Hamdy, P. Jelena Mitrovi, and M. Granitzer, "Neural Models for Offensive Language Detection Masterarbeit von," 2021.

[22] and M. R. Y. H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, "Text Mining in Big Data Analytics," *Big Data Cogn. Comput.*, vol. 4, no. 1, 2020, doi. https.//doi.org/10.3390/bdcc4010001.

[23] "Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles." Accessed. Dec. 22, 2024. [Online]. Available. https.//www.researchgate.net/publication/323387293_Unsupervised_Content-Based_Identification_of_Fake_News_Articles_with_Tensor_Decomposition_Ensembles

[24] N. Oswal, "Identifying and Categorizing Offensive Language in Social Media," *arXiv.2104.04871*, 2021, doi. https.//doi.org/10.48550/arXiv.2104.04871.

[25] "Natural language Processing Based Fake News Detection using Text Content Analysis with LSTM - Peer-reviewed Journal." Accessed. Dec. 22, 2024. [Online]. Available. https.//ijarcce.com/papers/natural-language-processing-based-fake-news-detection-using-text-content-analysis-with-lstm/

[26] and J. H. D. S. Kaddoura, G. Chandrasekaran, D. E. Popescu, "A systematic literature review on spam content detection and classification," *PeerJ Comput. Sci.*, vol. 8, p. 830, 2022, doi. 10.7717/peerj-cs.830.

[27] and M. K. N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, "Spam Review Detection Techniques. A Systematic Literature Review," *Appl. Sci.*, vol. 9, no. 5, p. 987, 2019, doi. https.//doi.org/10.3390/app9050987.

[28] and D. S. Prabhjot Kaur, Rajdavinder Singh Boparai, "Hybrid Text Classification Method for Fake News Detection," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 5, 2019, [Online]. Available. https.//www.ijeat.org/wp-content/uploads/papers/v8i5/E7622068519.pdf

[29] H. L. Xin Lyu, Yuxian Gu, Xu Han, "Adapting Meta Knowledge Graph Information for Multi-Hop Reasoning over Few-Shot Relations," 2019, [Online]. Available. https.//www.researchgate.net/publication/335564869_Adapting_Meta_Knowledge_Grap

h_Information_for_Multi-Hop_Reasoning_over_Few-Shot_Relations

[30] and S.-F. C. D. Lu, S. Whitehead, L. Huang, H. Ji, "Entity-aware Image Caption Generation," *Proc. 2021 Conf. Empir. Methods Nat. Lang. Process.*, 2018, doi. 10.18653/v1/D18-1435.

[31] T. He and J. Glass, "Negative Training for Neural Dialogue Response Generation," *Assoc. Comput. Linguist. Conf.*, 2019, doi. 10.18653/v1/2020.acl-main.185.

[32] D. R. Siyi Liu, Sihao Chen, Xander Uyttendaele, "MultiOpEd. A Corpus of Multi-Perspective News Editorials," *Proc. 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, pp. 4345–4361, 2021, [Online]. Available. https.//aclanthology.org/2021.naacl-main.344/

[33] G. Singh, I. J. Marshall, J. Thomas, J. Shawe-Taylor, and B. C. Wallace, "A neural candidate-selector architecture for automatic structured clinical text annotation," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. Part F131841, pp. 1519–1528, Nov. 2017, doi. 10.1145/3132847.3132989.

[34] and H. L. K. Shu, D. Mahudeswaran, S. Wang, "Hierarchical Propagation Networks for Fake News Detection. Investigation and Exploitation," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 14, 2020, doi. https.//doi.org/10.1609/icwsm.v14i1.7329.

[35] and S. L. J. Alghamdi, Y. Lin, "A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection," *Information*, vol. 13, no. 12, p. 576, 2022, doi. https.//doi.org/10.3390/info13120576.

[36] J. L. Takeshi Kurashima, Tim Althoff, "Modeling Interdependent and Periodic Real-World Action Sequences," *ACM Digit. Libr.*, pp. 803–812, 2018, doi. https.//doi.org/10.1145/3178876.3186161.

[37] H. G. Hanming Deng, Yang Hua, Tao Song, Zhengui Xue, Ruhui Ma, Neil Robertson, "Reinforcing Neural Network Stability with Attractor Dynamics," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, doi. https.//doi.org/10.1609/aaai.v34i04.5787.

[38] C. Z. Nora Hollenstein, "Entity Recognition at First Sight. Improving NER with Eye Movement Information," *Assoc. Comput. Linguist. Conf.*, 2019, [Online]. Available. https.//aclanthology.org/N19-1001/

[39] N. Cao, S. Ji, D. K. W. Chiu, and M. Gong, "A deceptive reviews detection model. Separated training of multi-feature learning and classification," *Expert Syst. Appl.*, vol. 187, p. 115977, Jan. 2022, doi. 10.1016/J.ESWA.2021.115977.

[40] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis. A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, p. e1253, Jul. 2018, doi. 10.1002/WIDM.1253.

[41] M. Kim, D. A. McFarland, and J. Leskovec, "Modeling afiinity based popularity dynamics," *Int. Conf. Inf. Knowl. Manag. Proc.*, vol. Part F131841, pp. 477–486, Nov. 2017, doi. 10.1145/3132847.3132923.

[42] "Logistic Boosted Algorithms for Securing Smart Homes Against Anomalies and Security Attacks." Accessed. Dec. 22, 2024. [Online]. Available. https.//www.researchgate.net/publication/380375279_Logistic_Boosted_Algorithms_for_Securing_Smart_Homes_Against_Anomalies_and_Security_Attacks

[43] "Detecting phishing e-mails using Text Mining and features analysis." Accessed. Dec. 22, 2024. [Online]. Available. https.//www.researchgate.net/publication/357053201_Detecting_phishing_e-mails_using_Text_Mining_and_features_analysis