A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

by

NAGOTHU RAMYA SREE **2203A52042**

Under the guidance of

Dr. D. RAMESH

Assistant Professor, School of CS&AI.

SR University, Ananthsagar,Warangal,Telagnana-506371

CONTENTS

CHAPTER 1

DATASET

Project -1

The **IDS Intrusion CSV dataset** on Kaggle is a labeled collection of network traffic data used for detecting potential intrusions in cybersecurity systems. It contains features representing various aspects of network connections, such as protocol type, source and destination ports, duration, byte counts, and flag status. The key purpose of this dataset is to enable the development and evaluation of **Intrusion Detection Systems (IDS)** by training models to distinguish between normal and malicious activities, such as DoS attacks, brute force attempts, and port scanning. This makes it a valuable resource for machine learning and cybersecurity research.

Project – 2

The **Satellite Image Classification** dataset by Mahmoud Reda on Kaggle comprises satellite images categorized into four classes: **Cloudy**, **Desert**, **Green Area**, and **Water**. This dataset is designed to facilitate the training and evaluation of deep learning models, particularly Convolutional Neural Networks (CNNs), for land cover classification tasks. Researchers and developers have utilized this dataset to build models achieving high accuracy in distinguishing various terrain types, making it a valuable resource for applications in remote sensing, environmental monitoring, and geospatial analysis.

Project – 3

The **News Category Dataset** by Rishabh Misra is a comprehensive collection of approximately 210,000 news headlines sourced from HuffPost, spanning from 2012 to 2022. Each entry in the dataset includes the article's headline, short description, publication date, authors, and a category label, encompassing 42 distinct categories such as Politics, Entertainment, Wellness, and Business. This dataset is particularly valuable for natural language processing tasks like text classification, topic modeling, and sentiment analysis, as it provides a rich and diverse set of real-world news data. Its structured format and extensive coverage make it an excellent resource for training and evaluating machine learning models in various text analysis

METHODOLOGY

Project – 1

Dataset Preparation:

The Intrusion Detection dataset contains network traffic logs, including various numerical features and attack labels. Initially, the dataset was loaded from a CSV file, and null or infinite values were handled. Irrelevant or redundant columns were dropped to ensure the data was clean and focused on essential features for detecting intrusions.

Data Preprocessing:

Numerical features were standardized using StandardScaler to bring all values to a similar scale, which improves model performance. The target labels (types of network traffic: attack or benign) were encoded using LabelEncoder for compatibility with machine learning algorithms.

Feature Selection:

Only the most relevant numerical features were selected based on initial analysis to reduce complexity and enhance model efficiency.

Model Training:

Different machine learning models like Logistic Regression, Gradient Boosting, and SVM were trained on the processed dataset. These models were evaluated based on accuracy, precision, recall, and F1-score.

Performance Evaluation:

Confusion matrices and metric scores were used to analyze how well each model detected network intrusions. Ensemble models such as Gradient Boosting showed balanced and effective results in identifying attacks.

Project -2

Dataset:

Labeled satellite images representing land cover types (e.g., buildings, forests, roads, water bodies) for classification.

Preprocessing:

Images resized to 64x64 pixels, normalized between 0-1, with augmentation techniques like rotation, flipping, and zooming applied.

Model Architecture:

CNN with convolutional, max-pooling, and dropout layers to extract features and reduce overfitting.

Training:

Categorical cross-entropy loss, with a separate validation set for performance monitoring.

Evaluation Metrics:

Accuracy, confusion matrix, and classification reports (precision, recall, F1-score) to assess model performance.

Project – 3

Dataset Preparation:

The dataset includes news articles labeled across multiple categories such as politics, sports, technology, and business.

Preprocessing:

All text samples were cleaned by removing punctuation, converting to lowercase, removing stopwords, and tokenizing. Sequences were padded to ensure consistent input length.

Feature Extraction:

Text was converted into sequences using Keras Tokenizer and embedded into dense vectors to capture word semantics.

Model Architecture:

A deep LSTM model was developed to learn temporal patterns in the text. The model includes dropout layers to reduce overfitting and ends with a softmax layer for multiclass prediction.

Model Training:

The model was trained using categorical cross-entropy loss and Adam optimizer, with early stopping applied to prevent overfitting.

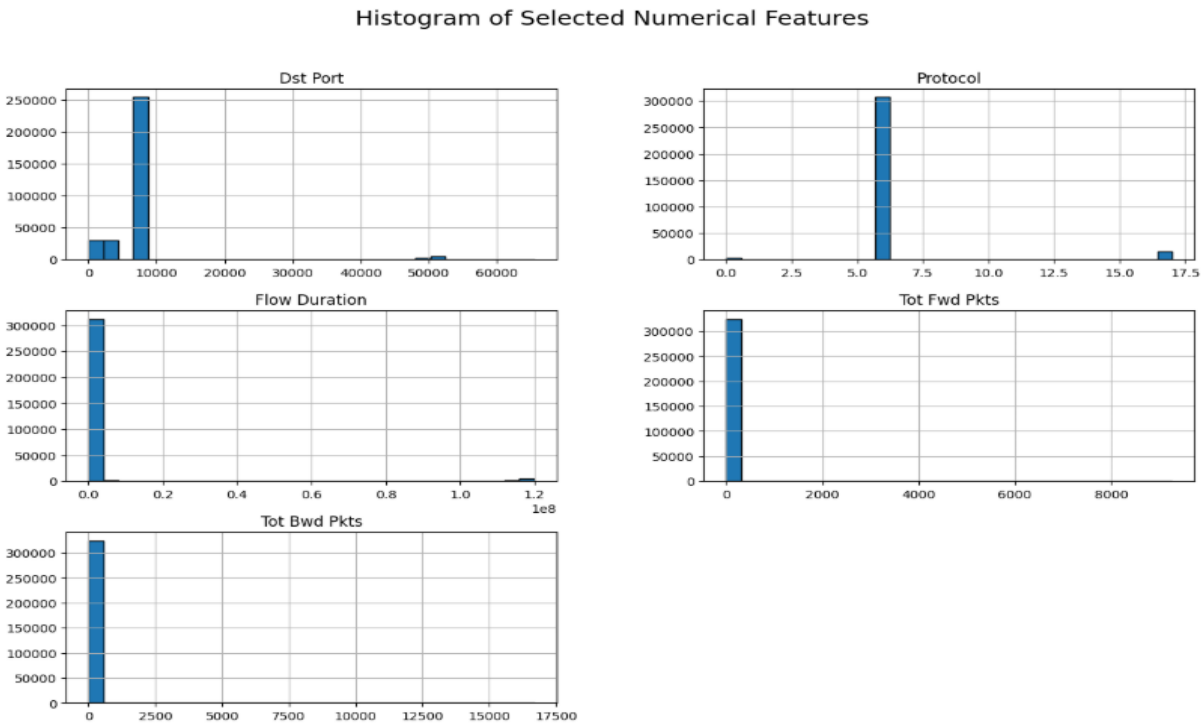
Performance Evaluation:

Model performance was evaluated using accuracy, precision, recall, and F1-score across categories. Confusion matrix and learning curves were used for visual analysis.

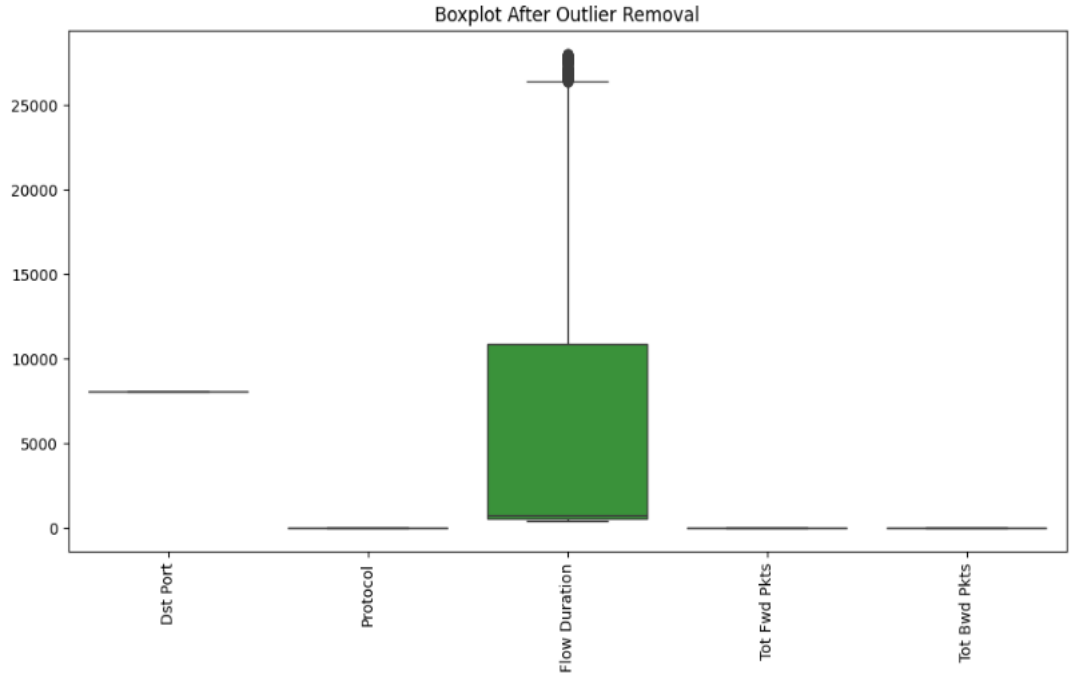
CHAPTER – 3

RESULTS

Project-1



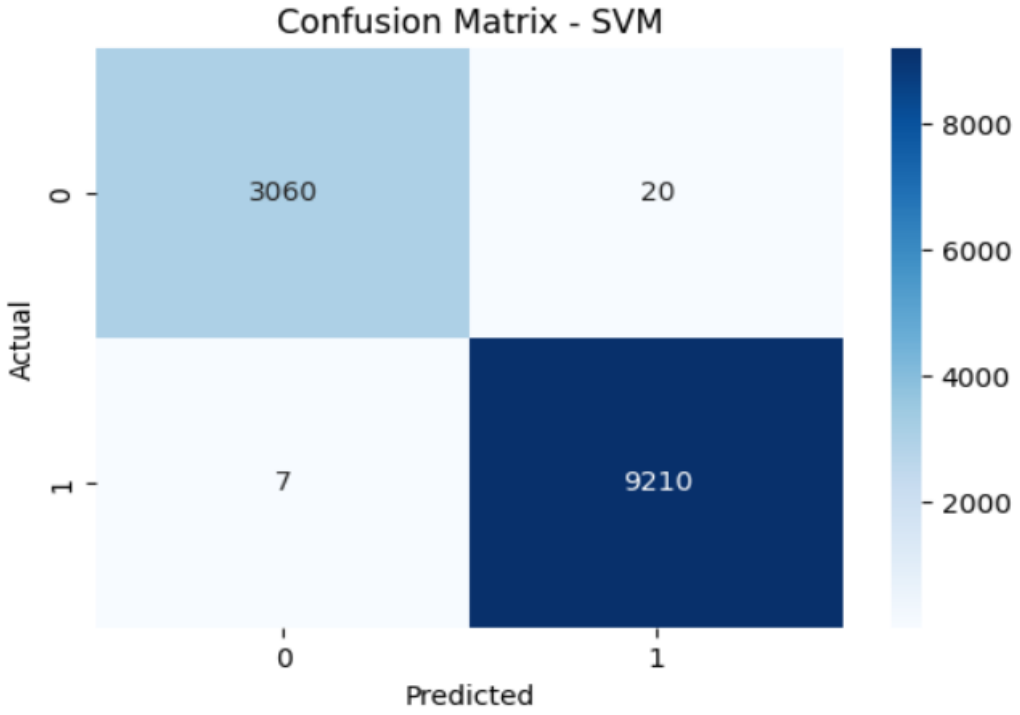
Box Plot



Classification Report

	Accuracy	Precision	Recall	F1-score
Random Forest	0.999919	0.999919	0.999919	0.999919
Logistic Regression	0.995202	0.995219	0.995202	0.995189
SVM	0.997804	0.997804	0.997804	0.997803

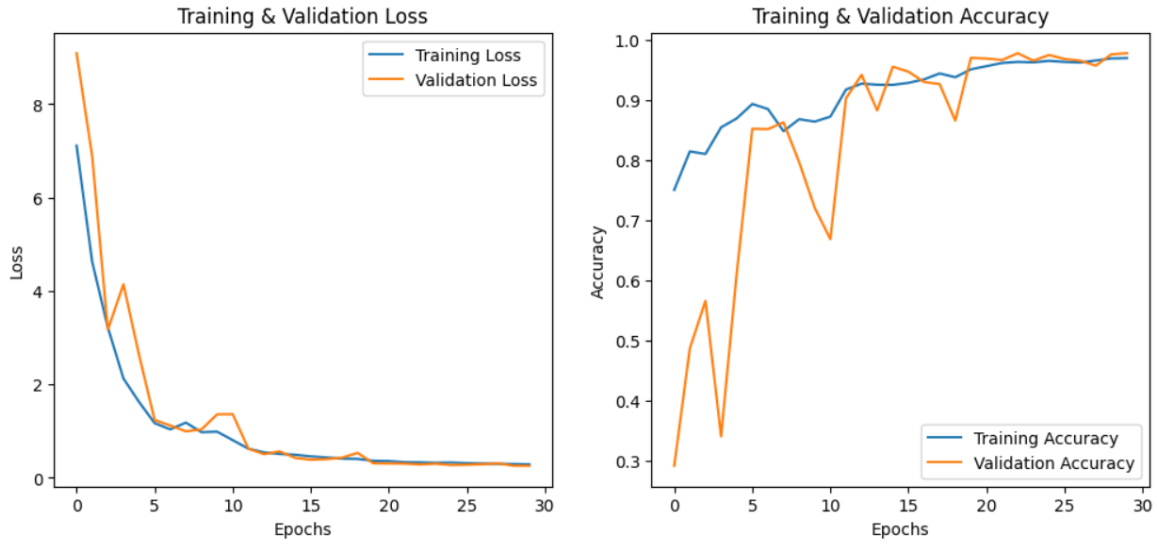
Random Forest model achieved the highest accuracy, precision, recall, and F1-score, all reaching approximately 99.99%, indicating almost perfect classification performance. The Support Vector Machine (SVM) also performed exceptionally well with around 99.78% across all evaluation metrics. Although slightly lower, Logistic Regression still delivered strong results with values around 99.52%



Project – 2

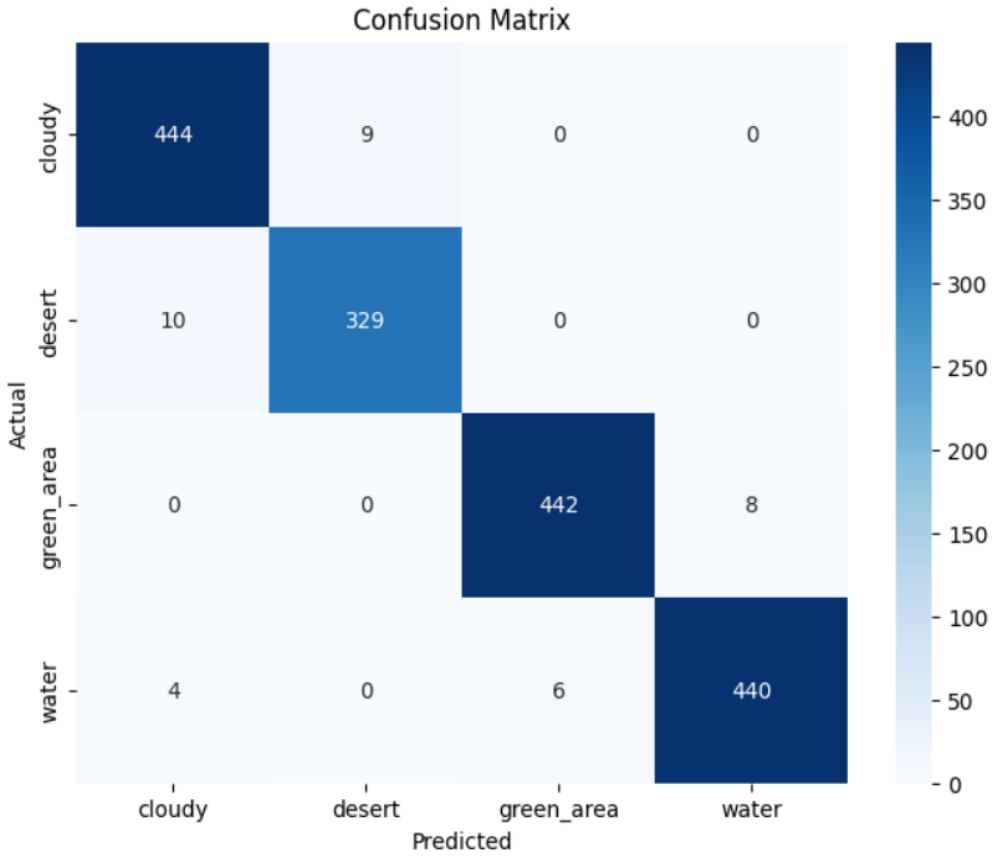
```
100%|██████████| 1510/1510 [00:47<00:00, 31.52it/s]
100%|██████████| 1131/1131 [00:37<00:00, 29.83it/s]
100%|██████████| 1500/1500 [00:51<00:00, 29.07it/s]
100%|██████████| 1500/1500 [00:54<00:00, 27.60it/s] Conversion to grayscale completed!
```

Showing each variety in its original RGB color and in grayscale at two different resolutions, as well as the original RGB resized. This standardized presentation is useful for various image processing and analysis tasks.



The first graph shows a steady decrease in both training and validation loss, indicating effective learning and minimal overfitting.

The second graph demonstrates a consistent rise in accuracy, with both training and validation accuracy converging near 98%, reflecting strong model performance.



The confusion matrix illustrates the model's strong performance in classifying satellite images across four land cover types: cloudy, desert, green_area, and water. Most predictions lie along the diagonal, indicating accurate classification. The model correctly identified 444 cloudy, 329 desert, 442 green_area, and 440 water images, with only a few misclassifications. Minor confusion is seen between cloudy and desert, as well as green_area and water, but overall, the model demonstrates high accuracy and reliability in distinguishing between different geographical features.

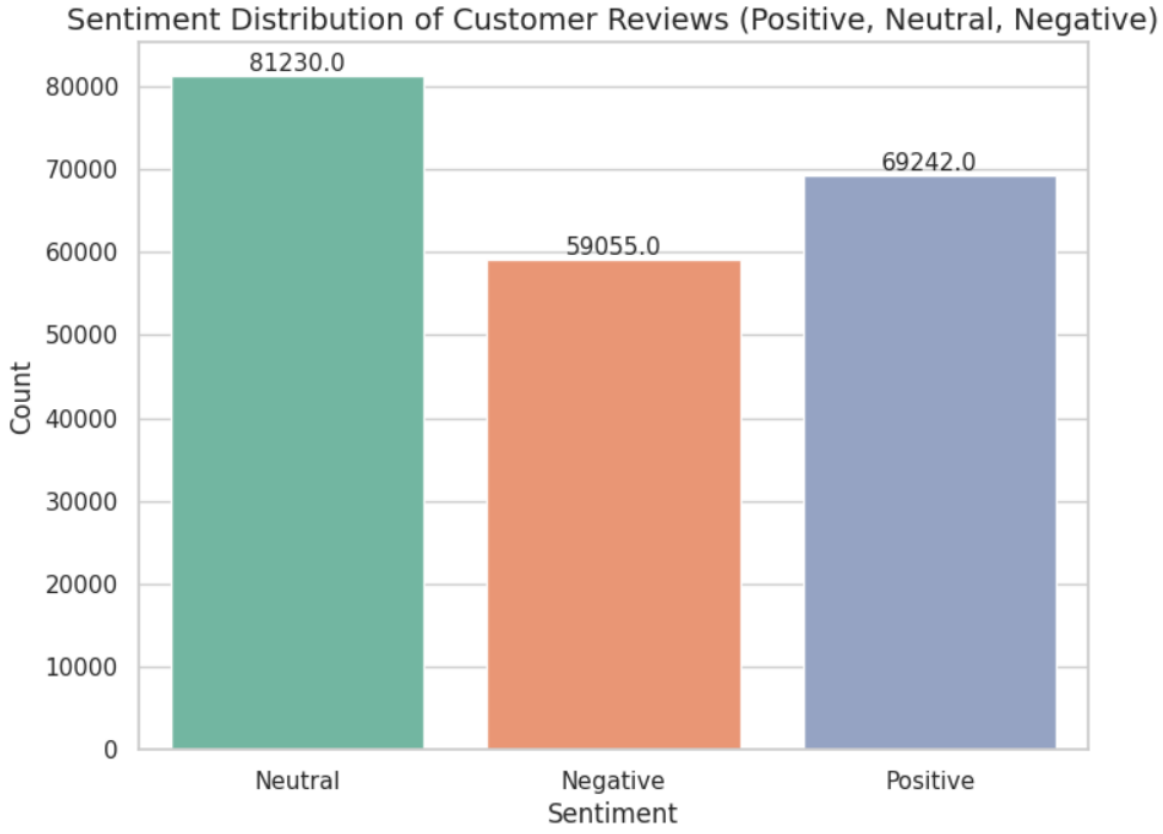
```
Z-score: -1.7212, P-value: 0.0852
T-test Statistic: 2.0701, P-value: 0.0461
ANOVA F-statistic: 4.2892, P-value: 0.0428
```

Z-test: Z-score = -1.7212, p-value = 0.0852 → Not significant ($p > 0.05$), so the null hypothesis is not rejected.

T-test: T-statistic = 2.0701, p-value = 0.0461 → Statistically significant ($p < 0.05$), indicating a meaningful difference.

ANOVA: F-statistic = 4.2892, p-value = 0.0428 → Statistically significant ($p < 0.05$), suggesting variation among group means.

Project-3

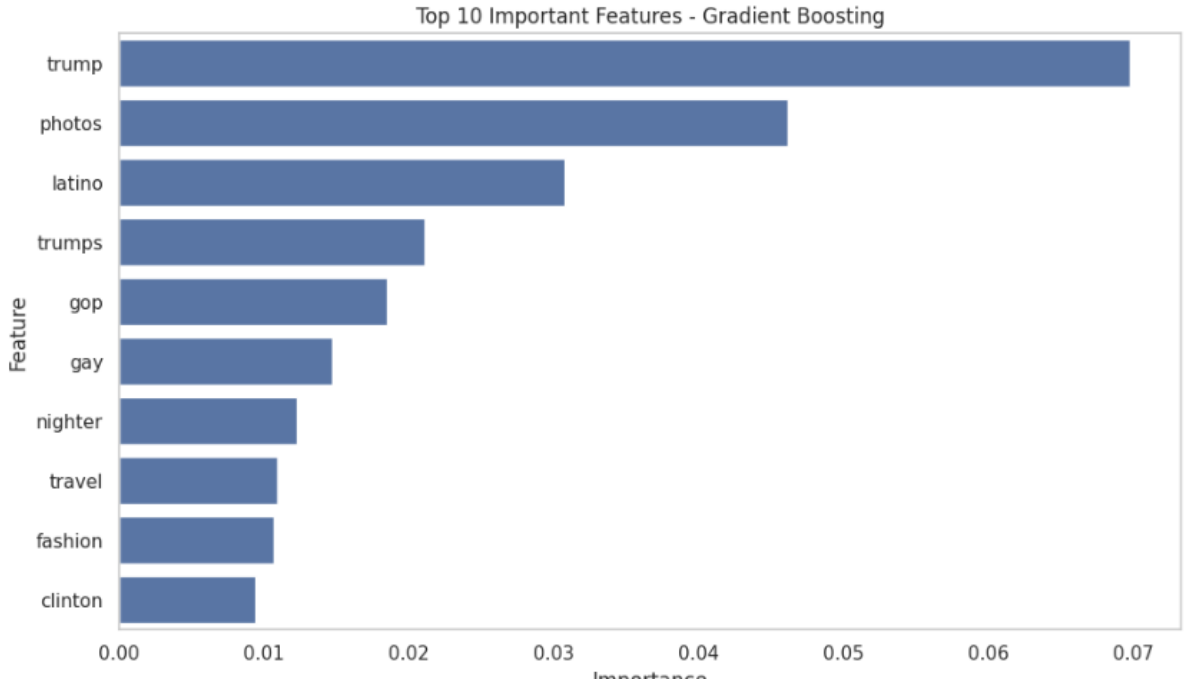


This bar chart illustrates the sentiment distribution of customer reviews across three categories: **Neutral**, **Negative**, and **Positive**. The **Neutral** sentiment has the highest number of reviews, totaling **81,230**, followed by **Positive** reviews at **69,242**. **Negative** reviews are the fewest, with **59,055** entries. This indicates that most customer feedback tends to be neutral, with a relatively balanced number of positive and negative sentiments, showing a moderately satisfied customer.

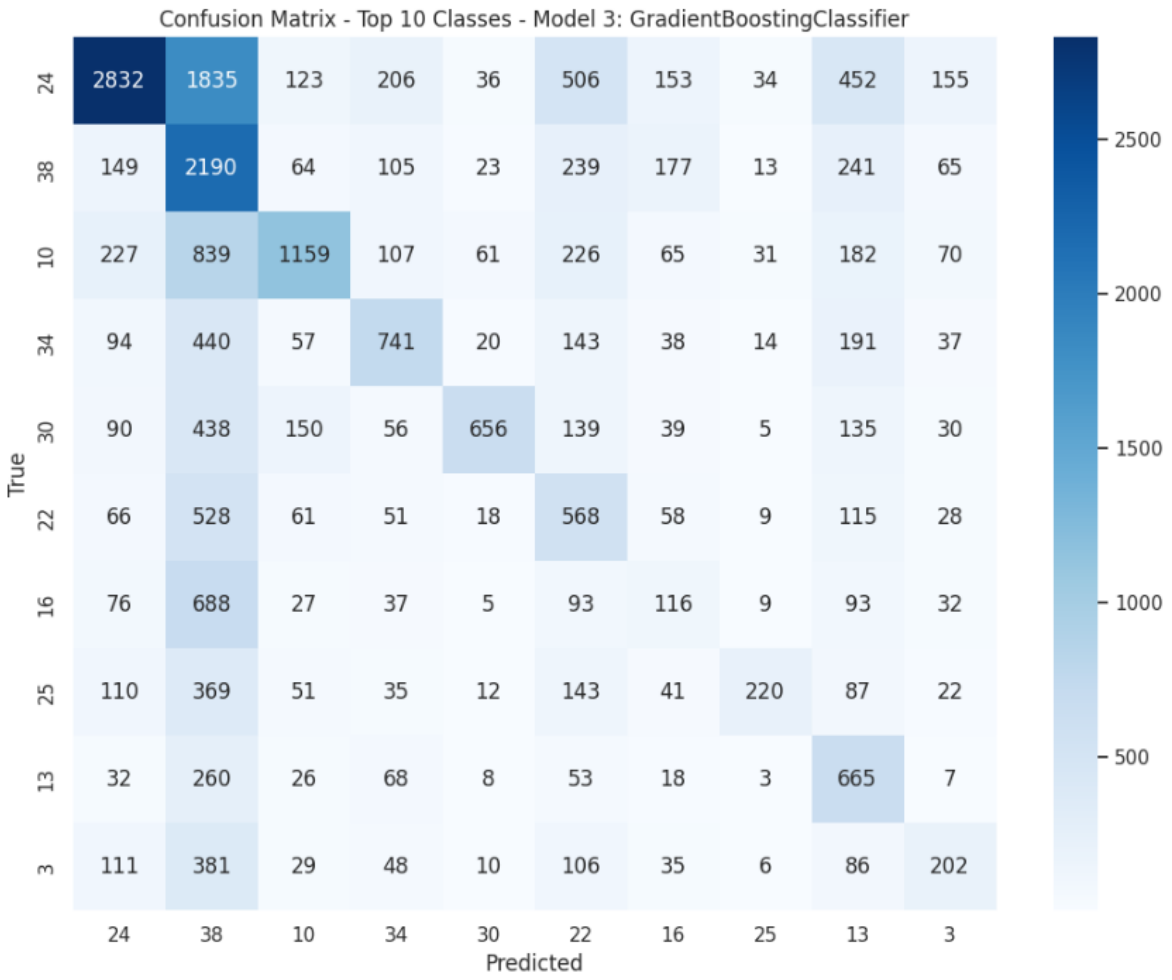
```
Training: Logistic Regression
Logistic Regression does not support feature importances.

Training: KNN
KNN does not support feature importances.

Training: Gradient Boosting
```



This image shows the top 10 most important words identified by a Gradient Boosting model. Words like "trump," "photos," and "latino" have the highest influence on predictions, indicating the dataset likely includes political and lifestyle content. The feature importance helps explain which terms most affect the model's decisions.



This confusion matrix visualizes the performance of a Gradient Boosting Classifier on the top 10 most frequent classes. Each row represents the actual class, and each column represents the predicted class. The diagonal values show correct predictions (e.g., class 24 and 38 have high correct counts), while off-diagonal values indicate misclassifications. Darker shades represent higher values, highlighting where the model performs well and where it confuses between classes.