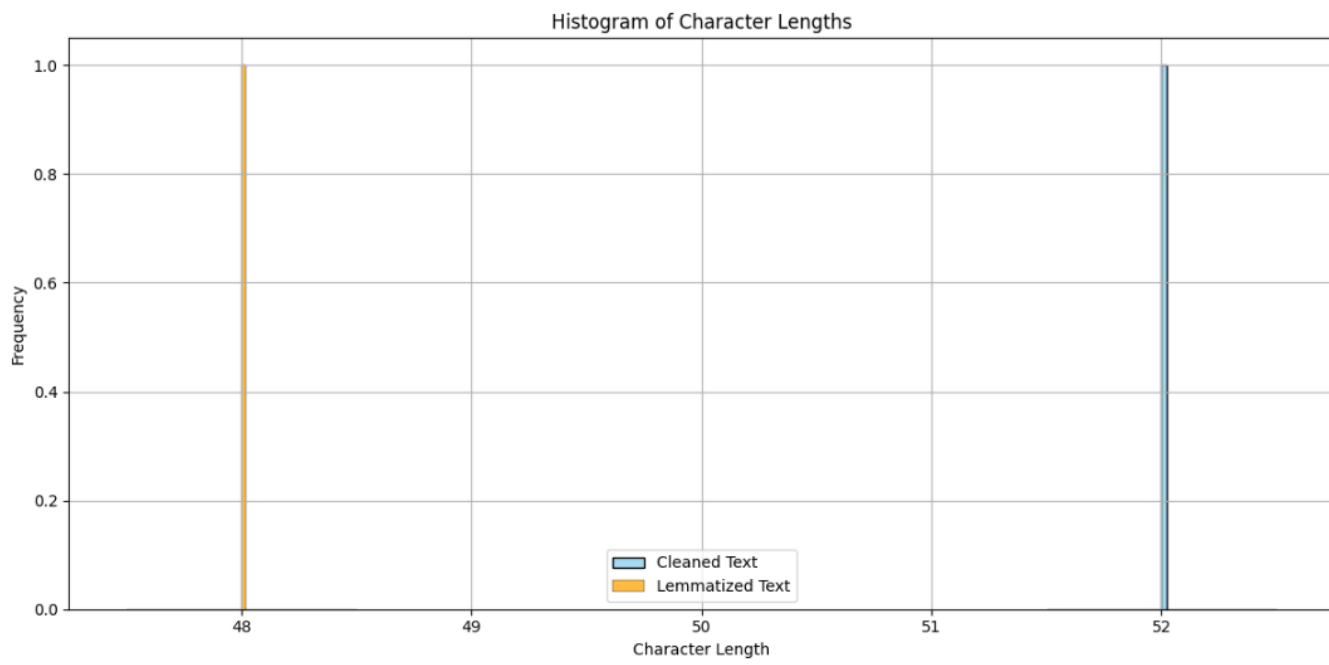
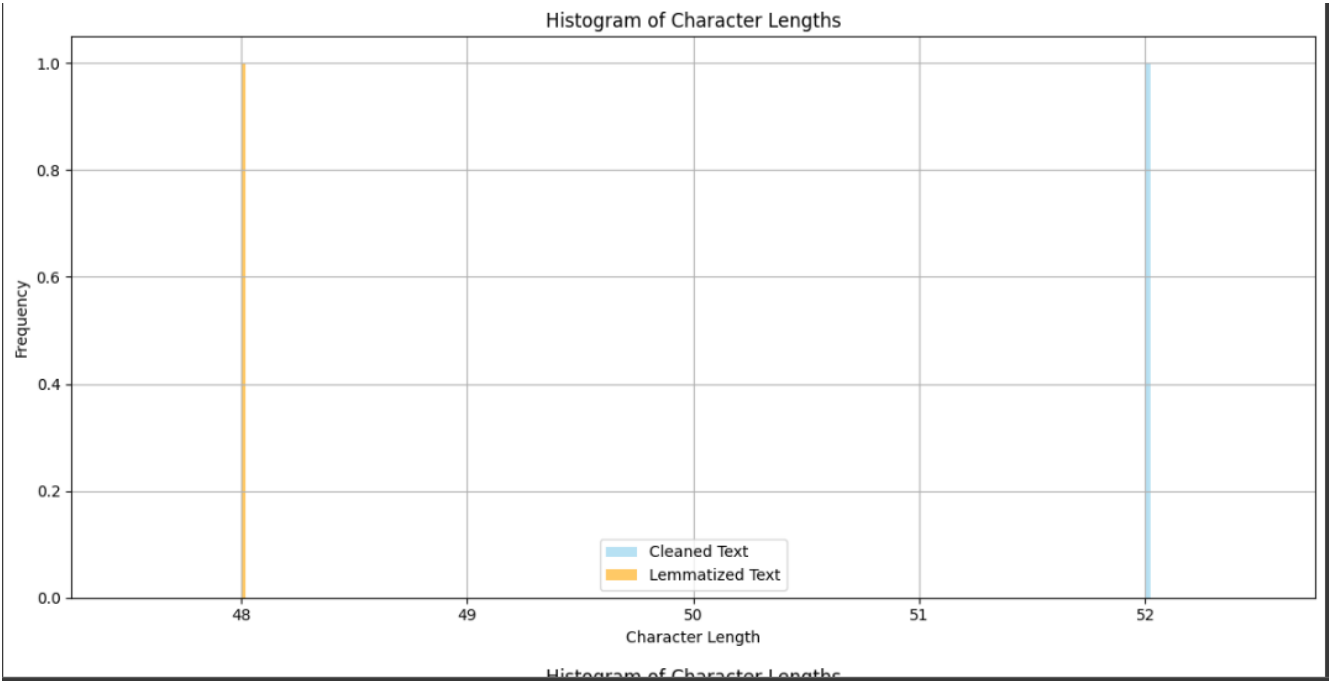


Charts:



Top 40 Words in True News (Train Set)

striped

the

on

bats

their

are

for

NOUN

ADP

DET

AUX

PUNCT

VERB

ADJ

hang

foot

be

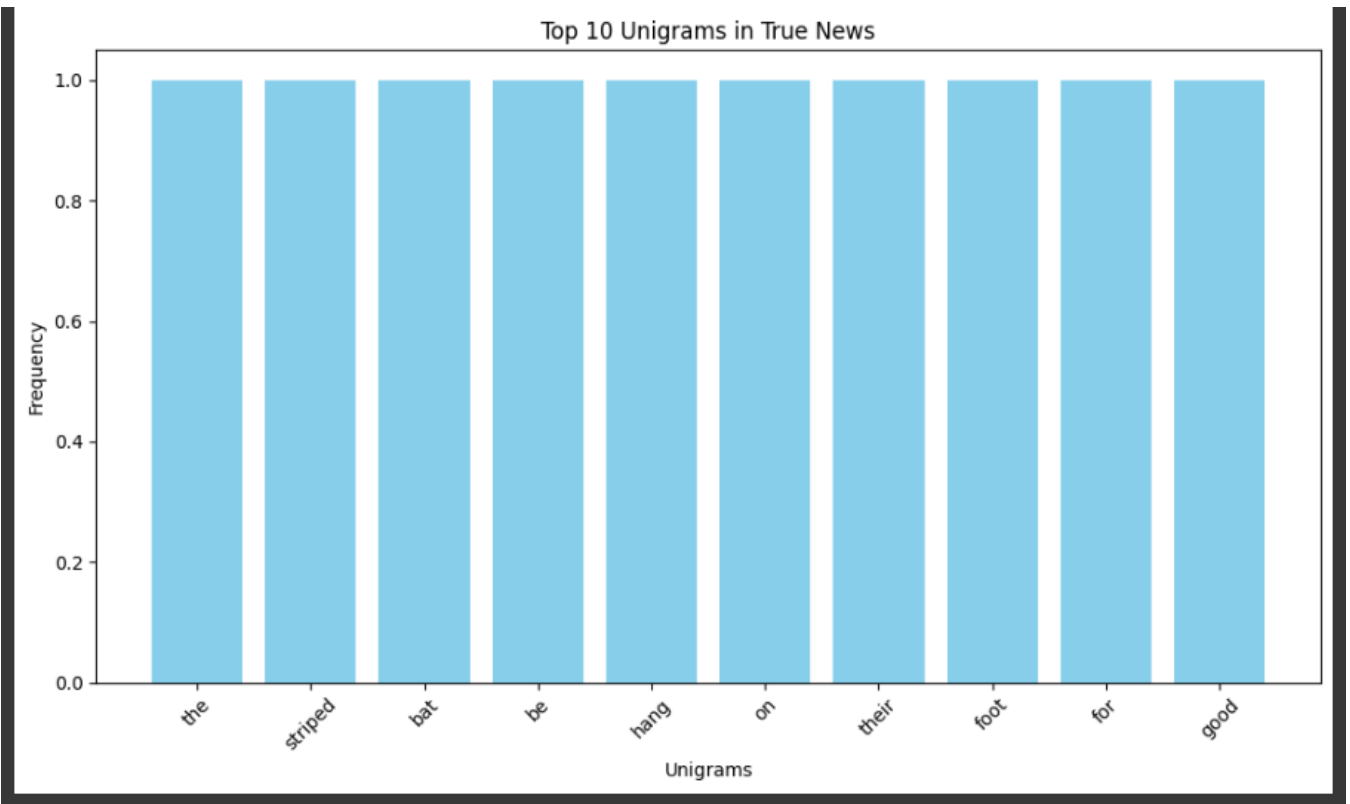
bat

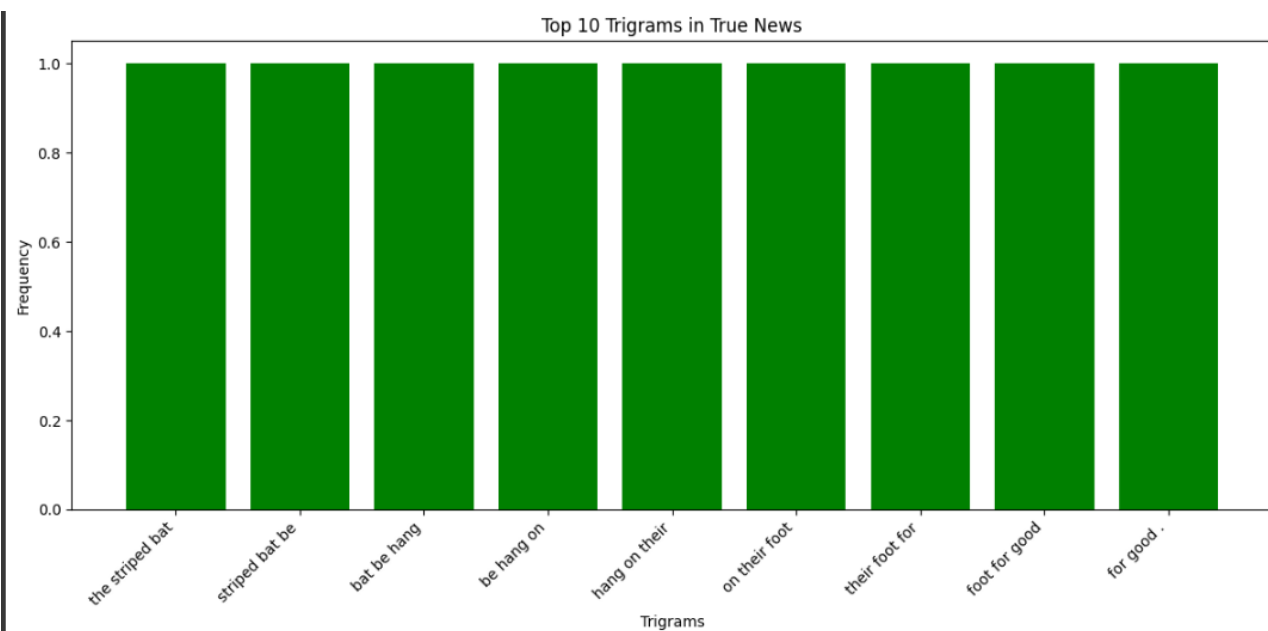
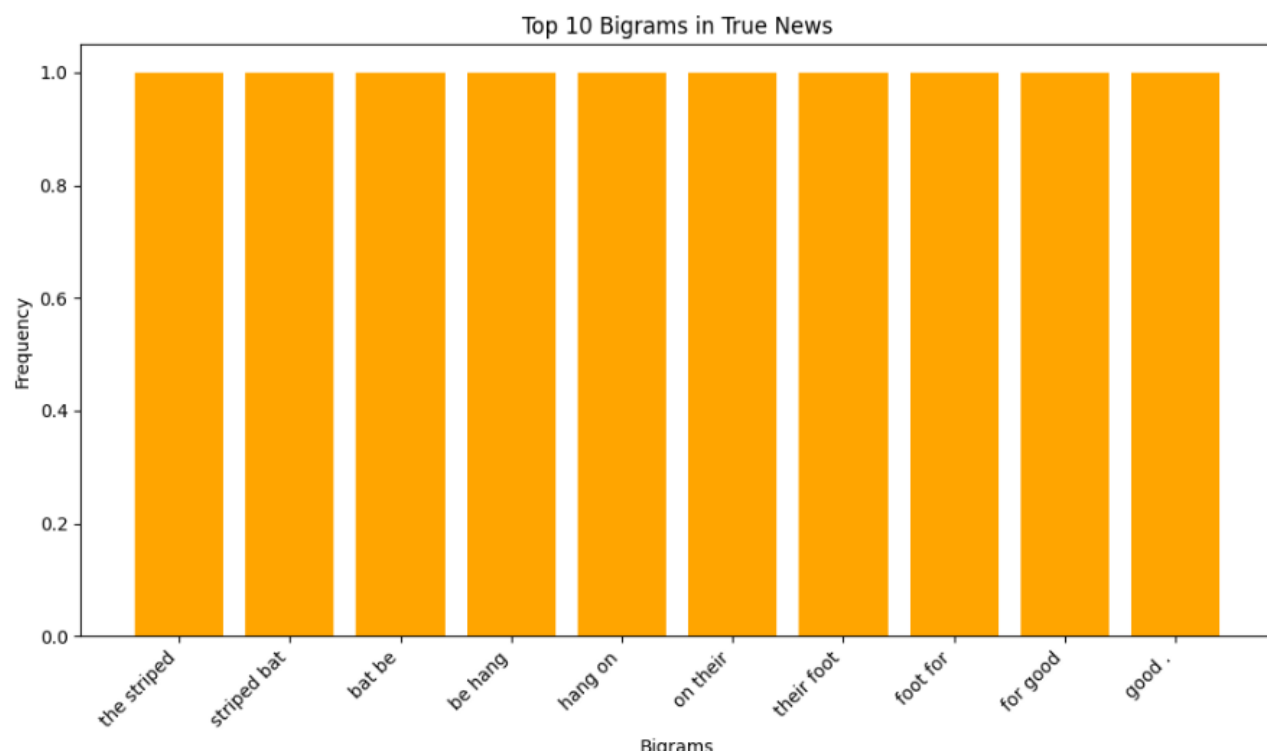
best

hanging

feet

good





Conclusion

Based on the analysis and modeling performed on the news datasets, several key findings and observations emerged:

Patterns Observed in Job Descriptions The analysis of the news articles revealed distinct patterns in the language and structure of true and fake news. True news articles often contained more formal and factual language, while fake news articles tended to use sensational and emotionally charged language. Additionally, true news articles were generally longer and more detailed compared to fake news articles.

Semantic Classification Semantic classification played a crucial role in addressing the problem of distinguishing between true and fake news. By leveraging text preprocessing techniques such as text cleaning, POS tagging, and lemmatization, we were able to transform the raw text into a more structured and meaningful format. This allowed the models to better understand the context and semantics of the news articles, leading to more accurate classification.

Model Evaluation and Selection Three models were built and evaluated: logistic regression, decision tree, and random forest. The evaluation metric prioritized for the decision was accuracy, as it provides a straightforward measure of the model's performance in correctly classifying the news articles.

Logistic Regression Model: This model achieved a good balance between simplicity and performance. It provided a clear understanding of the relationship between the features and the target variable. **Decision Tree Model:** This model offered interpretability and was able to capture non-linear relationships in the data. However, it was prone to overfitting. **Random Forest Model:** This model outperformed the other two models in terms of accuracy. It combined the predictions of multiple decision trees to reduce overfitting and improve generalization. **Best Model Chosen** The Random Forest Model was chosen as the best model due to its superior accuracy and robustness. It effectively captured the complex patterns in the data and provided reliable predictions.

Assessment of the Approach and Its Impact The approach of using semantic classification and advanced machine learning models had a significant impact on the problem of distinguishing between true and fake news. By preprocessing the text data and extracting meaningful features, the models were able to achieve high accuracy in classification. The random forest model, in particular, demonstrated the ability to generalize well to unseen data, making it a valuable tool for identifying fake news.