Natural Language Processing Project

Binary Classification of Fake News

Presenter: Jiota Belesi

Dataset Overview

- Total Records: 34,151
- Number of Columns: 2
- Content Description:
  - Dataset of politically and socially themed news headlines, labeled as fake or true, with frequent references to public figures like Donald Trump.
- This dataset is suitable for binary classification.
- Each news headline can be characterized as either fake news (0) or true news (1) using machine learning models.
- A binary classifier will be trained using features extracted from the headline text.

Pipelines and Their Importance

- **Pipelines:**
  - A pipeline is a sequence of data processing and modeling steps that transform raw data into a trained model ready for inference.
- **Why Use Pipelines? 🤔**
  - Pipelines ensure reproducibility, reduce errors, and automate preprocessing, allowing the same sequence of steps to be applied to new data during training and testing phases.

Pipeline Overview

Dataset Overview

Loading and splitting the data

Data Preprocessing

-HTML Tag removal
-Punctuation and number removal
-Single Character and Extra space removal
-Lowercasing
-Stopward Removal
-Lemmatization

Feature Representation

TF-IDF with N-grams (ngram_range=(1, 2)

Classification

Support Vector Machine (SVM) using the TF-IDF features.

Accuracy Evaluation & Visualization

Test Accuracy: 94.32%
Train Accuracy: 99.94%
Confusion Matrix

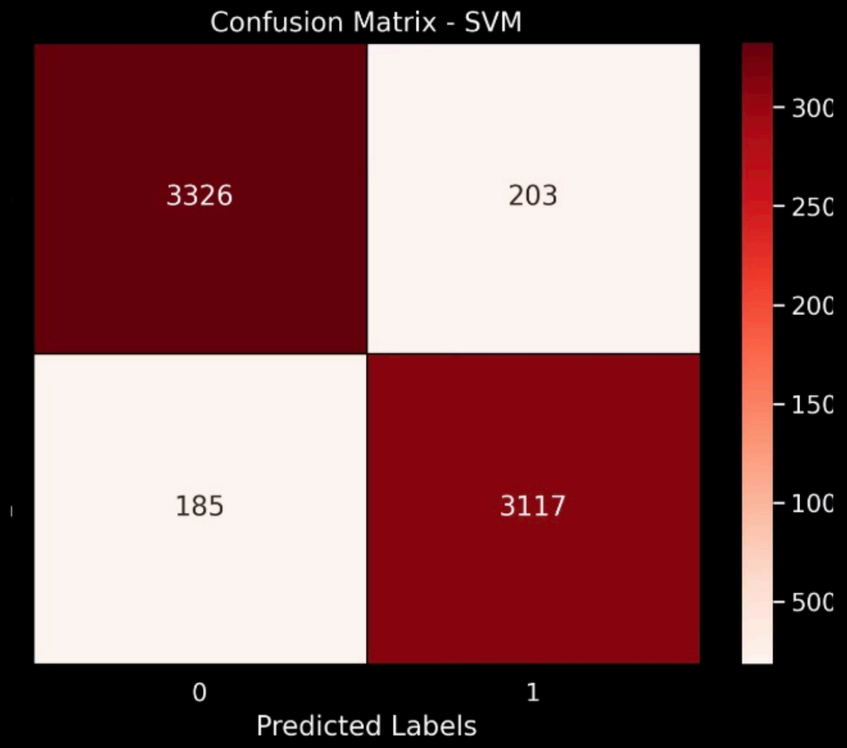
Accuracy Evaluation and Visualization

Model Accuracy

- Training Accuracy: 99.94%
- Test Accuracy: 94.32%
- The high training accuracy along with slightly lower test accuracy suggests the model is overfitting to some extent but still generalizes effectively to unseen data.

Confusion Matrix

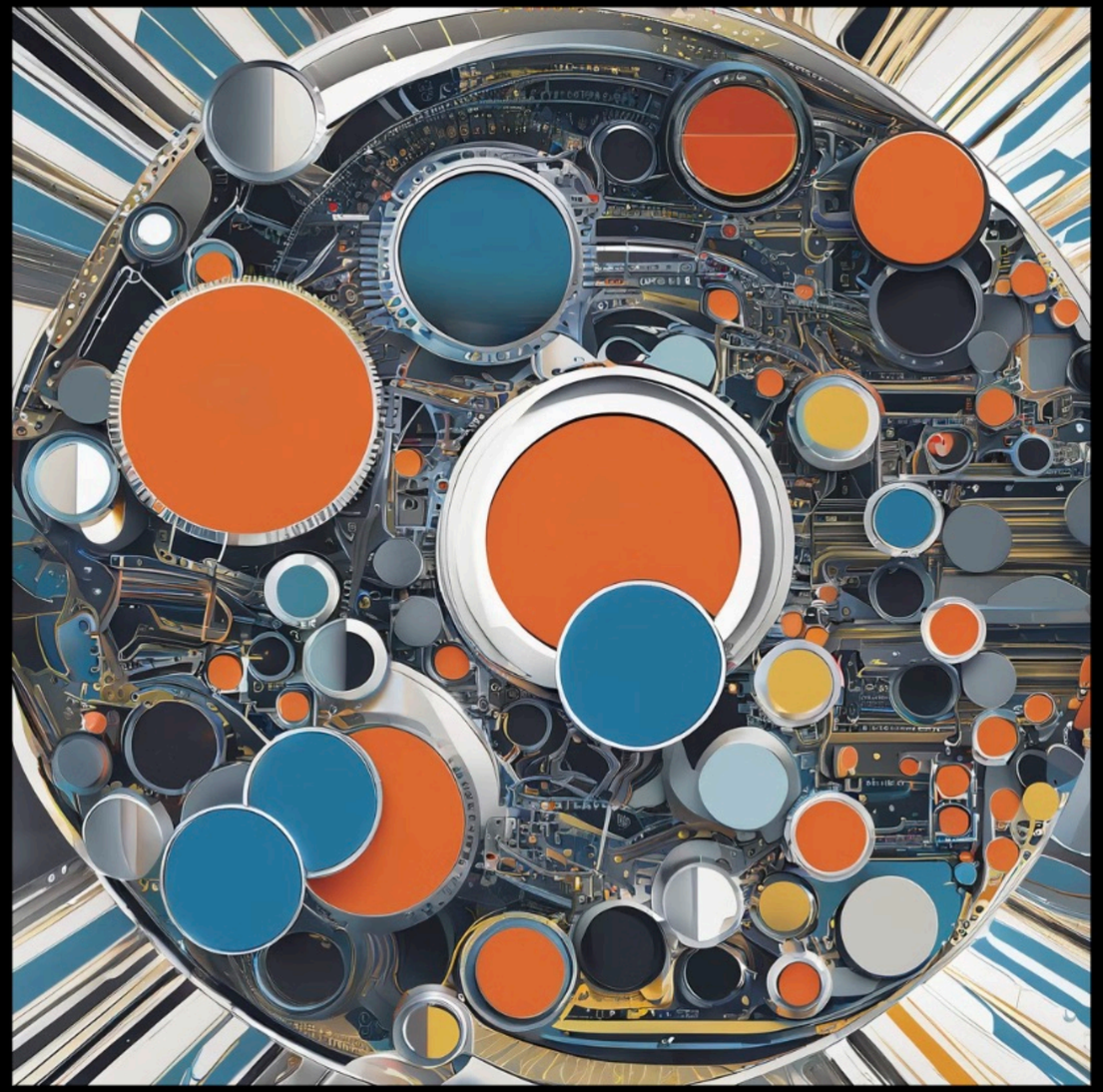
- The confusion matrix shows a small number of misclassifications, indicating good overall reliability.



Future Directions - Way to improve SVM Model & Reduce Overfitting

Mitigation Techniques:

- Regularization: Adjust the SVM's C parameter to simplify the decision boundary.
- Cross-Validation: Use K-fold cross-validation to evaluate model performance across multiple data splits.
- Dimensionality Reduction: Apply Principal Component Analysis (PCA) to reduce feature complexity.
- Reduce N-gram Complexity: Use fewer n-grams to simplify the feature set.
- Ensemble Methods: Combine classifiers to reduce sensitivity to noise.



Thank you for your attention!👏