FAKE NEWS DETECTION USING NLP

Project Summary:

This project aims to detect fake news articles by applying Natural Language Processing (NLP) techniques to classify news content as fake or true. The process involves several key steps:

🔹 1. Data Preparation

- Two datasets were used: one containing fake news articles and the other with true news.
- Title and text columns were combined into a single content field to consolidate information.

🔹 2. Text Preprocessing

- The raw text was cleaned by:
  - Converting to lowercase
  - Removing punctuation, digits, and special characters
  - Removing stopwords (e.g., "the", "is", "in")
  - Lemmatizing words (e.g., "running" → "run")

🔹 3. Tokenization and Lemmatization

- Tokenized the cleaned text into individual words.
- Applied lemmatization using NLTK's WordNetLemmatizer to reduce words to their base forms.

🔹 4. Feature Extraction using TF-IDF

- Transformed the preprocessed text into numerical features using TF-IDF Vectorization.
- Limited to the top 5000 most important words to reduce dimensionality.

🔹 5. Model Input Preparation

- Split the data into training and test sets.
- Converted the token lists back into strings for vectorization.
- Applied TfidfVectorizer to generate feature matrices for machine learning models.

✅ Outcome

The pipeline successfully transforms raw news data into clean, structured features ready for machine learning-based classification, helping to distinguish fake news from real news.

Data Analysis Key Findings

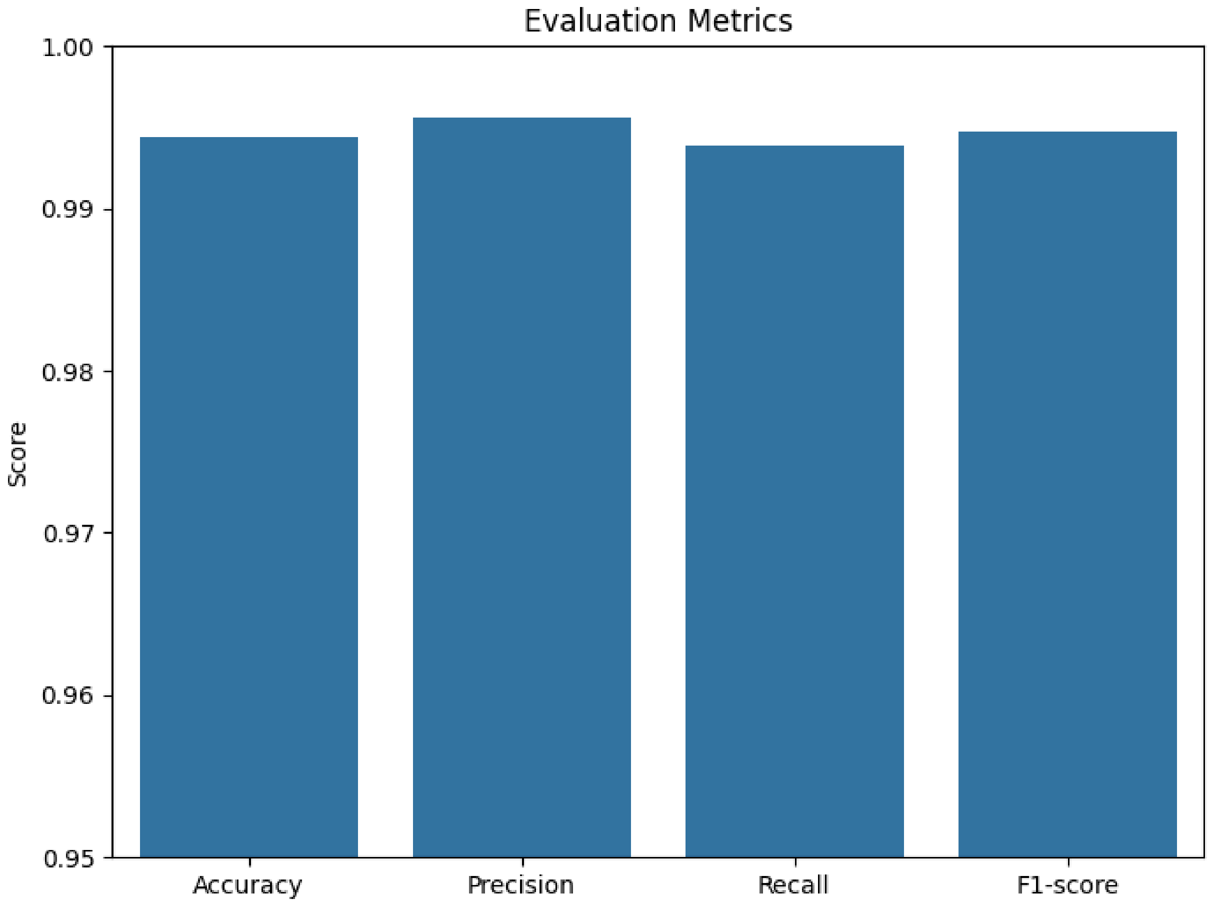
- The fake news dataset (Fake.csv) contains 23,481 articles, while the true news dataset (True.csv) contains 21,417 articles.
- Both datasets have 'title', 'text', 'subject', and 'date' columns, and neither contains missing values.
- The 'subject' column has 6 unique values in the fake news dataset and only 2 unique values in the true news dataset.
- After preprocessing, the text data was tokenized and lemmatized.

- The combined dataset was split into an 80% training set (35,918 samples) and a 20% testing set (8,980 samples).
- The text data was vectorized using TF-IDF with a maximum of 5000 features, resulting in training and testing matrices of shape (35918, 5000) and (8980, 5000) respectively.
- A Support Vector Classifier (SVC) with a linear kernel was trained on the vectorized training data.
- The trained model achieved the following performance metrics on the test set:
    - Accuracy: 0.9944
    - Precision: 0.9956
    - Recall: 0.9939
    - F1-score: 0.9947

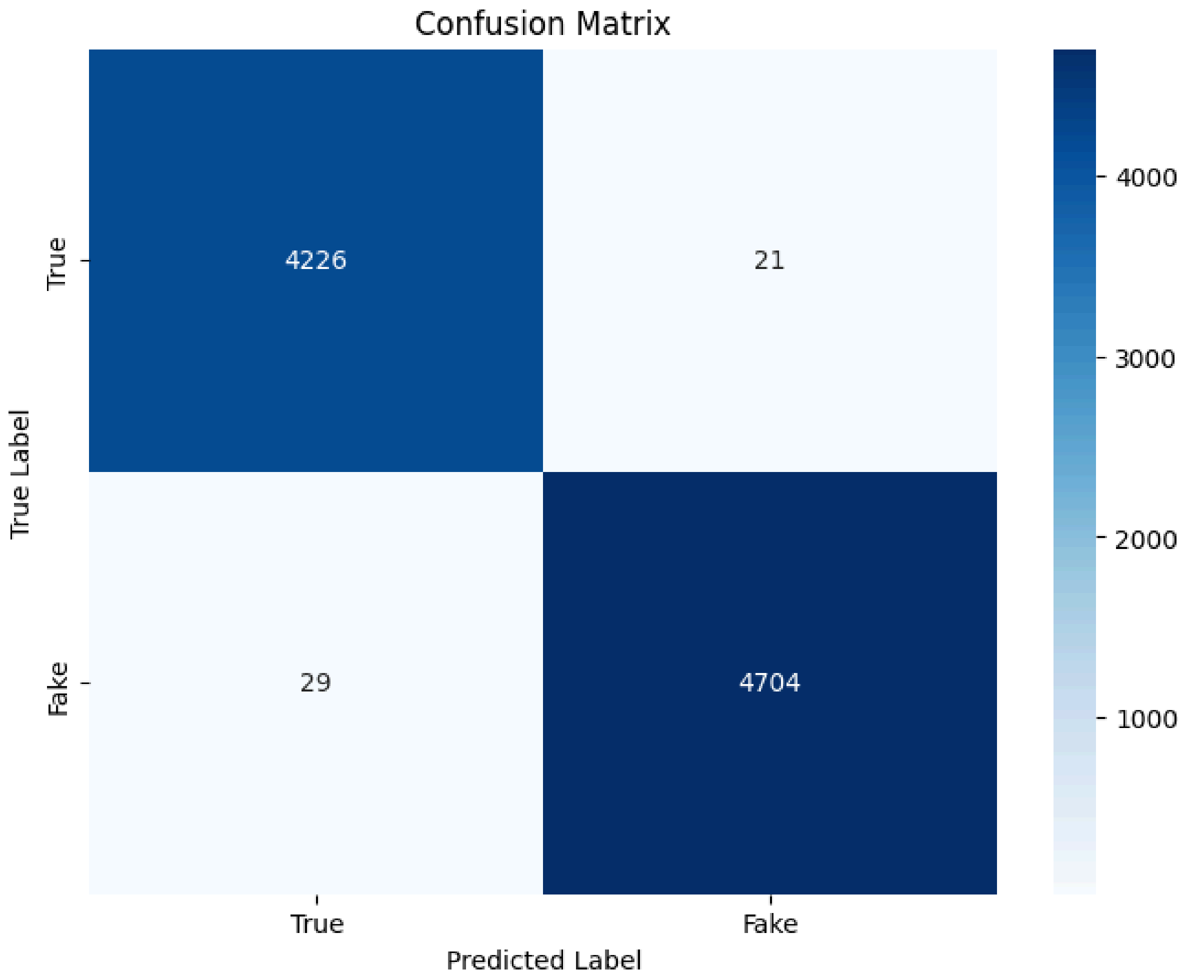
Insights or Next Steps

- The trained SVC model demonstrates excellent performance in classifying news articles as fake or real based on the provided datasets and preprocessing steps.
- Further analysis could explore the most important features (words) that contribute to the classification to gain insights into the linguistic patterns that differentiate fake and true news in these datasets.

Visualizations



All metrics are very high, ranging between **0.993 and 0.996**, indicating excellent model performance. The chart suggests the model is well-balanced and performs consistently across all evaluation metrics.



This image is a confusion matrix visualizing the performance of a classification model:

- **True Positives** (Fake correctly predicted as Fake): 4704
- **True Negatives** (True correctly predicted as True): 4226
- **False Positives** (True misclassified as Fake): 21
- **False Negatives** (Fake misclassified as True): 29

The model shows excellent classification performance, with very few misclassifications and strong balance between classes.