جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis 2

Text Analysis (Extra Model)

Dataset:

Spam Email

Student Name	ID
Alaa Khalil Bondagji	443000351
Jana Maher Jamal	443010542

Text Preprocessing:

Overview

For text preprocessing, we followed a systematic approach to clean and prepare the data for machine learning models. The key steps included

- **Data Cleaning**:
    - Removed URLs, punctuation, numbers, special characters, and stop words from the text.
    - Converted all text to lowercase and handled contractions such as "isn't" to "is not."
    - Created a `cleaned_text` column where the original reviews were cleaned for further processing.
    - Applied tokenization and vectorization to convert the text into numerical features using `CountVectorizer` and `TfidfVectorizer` to transform the text data.

- **Stop Words Removal**:
We removed common stop words and then further removed the most frequent words to reduce noise in the dataset, creating the `no_sw` and `wo_stopfreq` columns.

- **Lemmatization**:
After removing frequent words, lemmatization was applied to bring words to their base form, improving the quality of text analysis.

Model Performance:

Three Naive Bayes models were trained and evaluated on the processed text data: Complement Naive Bayes (CNB), Multinomial Naive Bayes (MNB), and Bernoulli Naive Bayes (BNB). The dataset was split into training and testing sets (80% training, 20% testing), and accuracy was calculated for each model.

- **Complement Naive Bayes (CNB):**
    - **Accuracy**: 95.33%

    - **Insights**: The highest accuracy among the models, making CNB the best choice for handling imbalanced text classification, especially for spam detection.
- **Multinomial Naive Bayes (MNB):**
    - **Accuracy**: 91.39%
    - **Insights**: Performed well, but slightly less accurate compared to CNB. MNB is still effective for standard text classification tasks.
- **Bernoulli Naive Bayes (BNB):**
    - **Accuracy**: 93.85%
    - **Insights**: BNB performed better than MNB but slightly lower than CNB, making it a solid option when dealing with binary features (spam vs. ham).

Insights Gained from the Analysis:

- **Spam vs. Ham**:
The dataset was highly imbalanced, with a large number of "ham" messages compared to "spam." CNB handled this imbalance the best, achieving the highest accuracy.

- **Important Words**:
By visualizing the most common words and performing further feature extraction, we were able to identify key terms commonly associated with spam and ham, which helped improve model performance.

- **Feature Extraction**:
Using both `CountVectorizer` and `TfidfVectorizer`, we extracted important n-grams and terms from the dataset, which were crucial in helping the models distinguish between spam and ham messages.

- **Confusion Matrices**:
The confusion matrices for all three models showed that CNB had the least misclassifications, particularly in identifying "spam" correctly.

- **Real-World Application**:
The insights from this analysis can be directly applied to email filtering systems, ensuring that spam messages are accurately detected while minimizing false positives on legitimate messages (ham). Additionally, the text preprocessing pipeline can be generalized to other NLP tasks.

Conclusion:

Through this analysis, we found that the Complement Naive Bayes model (CNB) provided the best performance, and the preprocessing steps (cleaning, vectorization, and lemmatization) significantly contributed to improving model accuracy. The results are practical for building robust spam detection systems.