Report on Email Spam Classification Using NLP Techniques

Introduction

The project's objective is to classify emails into "spam" or "ham" categories using machine learning and natural language processing (NLP) techniques. We have employed various preprocessing methods, feature extraction techniques, and a Naive Bayes classifier to achieve this goal.

Data Processing

Preprocessing

Here are the reasons for each preprocessing step:

1. **Lowercasing**: Ensures that the same words in different cases are treated identically, improving model consistency.

2. **URL Removal**: Prevents the model from being distracted by web addresses, which usually don't contribute to email content classification.

3. **HTML Tag Remova**l: Focuses analysis on actual content by removing HTML structure, irrelevant for understanding email text.

4. **Non-alphabetic Character Removal**: Simplifies text data by eliminating symbols and numbers that might not contribute to the overall meaning.

5. **Tokenization**: Splits text into manageable pieces or words, enabling systematic analysis and processing by the model.

6. **Stopword Removal**: Increases model efficiency by ignoring frequent but uninformative words, allowing focus on meaningful content.

7. **Lemmatization**: Enhances model accuracy by grouping various forms of a word into a single base form, capturing the essence of the word within the context.

Why we chose lemmatization

1. **Preservation of Semantic Meaning**: Lemmatization maintains the intrinsic meaning of words by reducing them to their dictionary forms, unlike stemming, which may lead to meaning loss.

2. **Contextual Awareness:** Lemmatization understands word context and part of speech, allowing it to accurately process words with multiple meanings, crucial for effective email classification.

3. **Reduction in Dimensionality**: It simplifies the feature space by consolidating variations of a word into its base form, aiding in the accuracy of machine learning models without sacrificing important textual information.

4**. Improved Model Performance**: By providing a more accurate representation of words, lemmatization contributes to more reliable features for classifiers, enhancing their ability to differentiate between spam and ham emails.

5. **Handling of Advanced Cases**: Lemmatization excels at processing irregular word forms through morphological analysis, ensuring correct reduction to root forms, which is advantageous for analyzing complex email content.

Exploratory Data Analysis (EDA) Section

Category	Word	Count	Avg Sentiment
Spam (Top by Count)	subject	1658	0.08
	com	993	0.11
	http	984	0.10
	company	921	0.09
	u	753	0.12
Ham (Top by Count)	ect	13897	0.02
	hou	7281	0.02
	enron	6555	0.01
	subject	6405	0.00
	deal	3534	0.02
All Emails (Top by Count)	ect	13908	0.02
	subject	8063	0.02
	hou	7289	0.02
	enron	6555	0.01
	com	3710	0.06
Spam (Top Negative Sentiment)	und	31	-0.04
	dosage	30	-0.04
	loading	27	-0.00
	expensive	27	-0.03
	ur	27	-0.13
Ham (Top Negative Sentiment)	hpl	2318	-0.04
	attached	1083	-0.05
	xl	1035	-0.11
	see	1020	-0.03
	nom	831	-0.07
All Emails (Top Negative Sentiment)	hpl	2318	-0.04
	see	1211	-0.01
	attached	1097	-0.05
	xl	1049	-0.10
	nom	832	-0.07

Most Frequent Words Analysis

Exploratory Data Analysis on the preprocessed email data revealed insightful patterns about the most frequent words in both spam and ham (non-spam) emails. By examining the lemmatized versions of these emails, several key observations were made.

Common Terms in Spam Emails: Words such as "company," "price," "information," and "email" were highly frequent in spam emails. These terms are indicative of promotional content, often focusing on offers, products, and services. The presence of words like "investment," "stock," and "million" further suggest financial incentives, a common theme in spam.

Common Terms in Ham Emails: In contrast, ham emails often contained words like "enron," "deal," "gas," and "meter," reflecting the business and operational nature of legitimate corporate communications. Terms like "please," "thanks," and "need" indicate polite requests and acknowledgments typical in professional exchanges.

Most Frequent Bigrams Analysis

Bigrams, or pairs of consecutive words, were also analyzed to understand common phrases within the dataset:

Spam Bigrams: Phrases like "http www" and "investment advice" were prevalent in spam emails, pointing to the inclusion of web links and financial recommendations. Other notable bigrams include "free pill" and "online pharmacy," which are characteristic of health-related spam.

Ham Bigrams: For ham emails, bigrams such as "hou ect" (referring to Houston, the location of Enron's headquarters) and "enron enron" underscored the company-specific language. Other frequent pairs like "please let" and "let know" reflect the collaborative nature of workplace communications.

Relevance to Prediction

The frequent words and bigrams offer a clear distinction between the language used in spam and ham emails, underscoring their relevance in classifying emails. Spam emails tend to focus on promotional language, financial opportunities, and unsolicited offers, while ham emails are characterized by corporate terminology, polite requests, and internal communications.

Sentiment Analysis of Common Words

The sentiment analysis of the most frequently occurring words in the dataset provides insights into the general tone and affective states present in the emails. In spam emails, words like "dosage," "loading," and "expensive" carry a mildly negative sentiment, potentially reflecting content typically found in unwanted commercial or promotional messages.

Conversely, in legitimate emails (ham), we observe words like "hpl" and "attached" with a slight negative sentiment, which may be related to the business context where issues or problems are discussed. Interestingly, the word "see" appears in both spam and ham emails with negative sentiment, suggesting it might be used in contexts requiring attention or action, which can have a negative connotation.

When looking at the most common words irrespective of sentiment, terms such as "subject," "ect" (possibly a shorthand or a misspelling of 'etc'), "hou," and "enron" dominate. The frequent occurrence of "enron" and associated terms like "hou" and "ect" in ham emails is expected, given that the dataset is sourced from Enron Corporation communication. Their presence helps in distinguishing between spam and ham, as they signal business-related conversations.

"Subject" being common across both spam and ham indicates its usage in email headers, making it a less distinctive feature. However, the average sentiment scores alongside these words may not provide significant discriminatory power, as they are relatively neutral or only slightly positive/negative.

Feature Engineering

Base Model Performance

The base model performance was established using a Bag-of-Words (BoW) approach with lemmatization:

Precision and Recall for 'Ham' (Label 0): High precision (0.9687) and recall (0.9330) indicate strong performance in identifying non-spam emails correctly.

Precision and Recall for 'Spam' (Label 1): Slightly lower precision (0.8605) and high recall (0.9320) suggest good spam detection but with some non-spam emails mistakenly classified as spam.

Effect of TF-IDF

Applying Term Frequency-Inverse Document Frequency (TF-IDF) did not improve the model performance significantly:

Accuracy: Dropped dramatically to 0.4548, showing a mismatch between the features extracted by TF-IDF and the classification task.

Precision and Recall: High precision (0.8036) but low recall (0.4548) indicate that while the model is confident in its spam predictions, it fails to identify a significant portion of spam emails.

Incorporating Sentiment Scores

Adding sentiment scores to the BoW features with lemmatization showed an improvement:

Accuracy: Increased to 0.9466, highlighting the value of sentiment analysis in distinguishing between spam and ham emails.

Overall Model Metrics: Precision (0.9538), recall (0.9466), and F1 score (0.9477) all indicate high performance, suggesting that sentiment provides meaningful contextual cues for classification.

Varying Vocabulary Sizes

Adjusting the vocabulary size in the BoW model produced nuanced differences:

Small Vocabulary (500 terms): Achieved an accuracy of 0.9296, demonstrating that even with fewer features, the model performs well.

Large Vocabulary (2000 terms): Slightly higher accuracy (0.9312) with marginally better precision and recall, indicating that increasing the number of features can capture more information but with diminishing returns.

Part-of-Speech (POS) Tagging

Using POS tagging as a feature exhibited a different aspect:

Accuracy: Was notably lower at 0.7347, suggesting that syntactic features alone are less effective for this specific task.

Precision and Recall: Lower scores (Precision: 0.7194, Recall: 0.7347) compared to other models indicate the challenges of relying solely on grammatical structures for spam classification.## Classification Experiments

Negation Handling Results

Accuracy: Maintained at 0.9466, the same level as observed when adding sentiment scores alone, suggesting that negation handling, in conjunction with lemmatization and BoW features, effectively contributes to the model's ability to accurately classify emails.

Precision, Recall, and F1 Score: With precision at 0.9538, recall at 0.9466, and F1 score at 0.9477, the model demonstrates high performance across these metrics, indicating a balanced and robust capability to differentiate between spam and ham emails.

Model Evaluation Results

The evaluation focused on various models combining Bag-of-Words (BoW), TF-IDF, sentiment analysis, negation handling, and Part-of-Speech (POS) tagging to classify emails as spam or ham. Each model was assessed based on accuracy, F1 score, precision, and recall for both classes (Class 0 for ham and Class 1 for spam) at different vocabulary sizes (2000, 1000, and 500).

Key Observations:

- BoW with Sentiment and POS at a vocabulary size of 2000 achieved the highest accuracy (0.9474) and F1 score (0.9480), indicating a strong balance of recall and precision. This model effectively leveraged syntactic and sentiment features to distinguish between spam and ham emails.

- Inclusion of Negation slightly reduced accuracy and F1 scores across vocabulary sizes when combined with sentiment and POS tagging. This suggests that while negation handling adds valuable context, its impact might be nuanced depending on the complexity and variability of the dataset.

- TF-IDF-based models generally showed a slight decrease in performance metrics compared to BoW models, particularly in the context of combining sentiment, negation, and POS features. This may indicate that the normalization of term frequencies in TF-IDF does not always capture the nuances necessary for spam detection as effectively as simple frequency counts.

- Vocabulary Size had a noticeable impact on model performance. Larger vocabulary sizes tended to yield better results, likely due to the increased ability to capture a broader range of informative features.

However, the gains from increasing vocabulary size diminished at higher levels, suggesting a point of diminishing returns.

Model Description	Vocabulary Size	Accuracy	F1 Score	Precision	Recall (Class 0)	Precision	Recall (Class 1)
BoW with Sentiment and POS	2000	0.9474	0.948	0.98	0.94	0.88	0.96
BoW with Sentiment, Negation, and POS	2000	0.9451	0.9458	0.98	0.94	0.87	0.96
BoW with Negation and POS	2000	0.9451	0.9458	0.98	0.94	0.87	0.96
TF-IDF with Sentiment and Negation	2000	0.9443	0.9449	0.98	0.94	0.88	0.95
BoW with Sentiment and Negation	2000	0.9443	0.945	0.98	0.94	0.87	0.96
TF-IDF with Sentiment and POS	2000	0.9404	0.941	0.97	0.94	0.88	0.94
TF-IDF with Sentiment, Negation, and POS	2000	0.9389	0.9391	0.96	0.95	0.89	0.91
TF-IDF with Negation and POS	2000	0.9389	0.9391	0.96	0.95	0.89	0.91
BoW with Sentiment and POS	1000	0.935	0.936	0.98	0.93	0.85	0.95
BoW with Sentiment, Negation, and POS	1000	0.9335	0.9344	0.97	0.93	0.85	0.94
BoW with Negation and POS	1000	0.9335	0.9344	0.97	0.93	0.85	0.94
TF-IDF with Sentiment and Negation	1000	0.9312	0.9317	0.96	0.94	0.87	0.92
TF-IDF with Sentiment and POS	1000	0.9304	0.931	0.96	0.94	0.86	0.92
TF-IDF with Sentiment, Negation, and POS	1000	0.9288	0.9292	0.96	0.94	0.87	0.9
TF-IDF with Negation and POS	1000	0.9288	0.9292	0.96	0.94	0.87	0.9
BoW with Sentiment, Negation, and POS	500	0.9281	0.9292	0.98	0.92	0.84	0.95
BoW with Negation and POS	500	0.9288	0.9299	0.98	0.92	0.84	0.95
BoW with Sentiment and POS	500	0.9273	0.9285	0.98	0.92	0.83	0.95
BoW with Sentiment and Negation	500	0.9227	0.9234	0.96	0.93	0.85	0.91
TF-IDF with Sentiment, Negation, and POS	500	0.9211	0.9212	0.94	0.94	0.87	0.87
TF-IDF with Negation and POS	500	0.9211	0.9212	0.94	0.94	0.87	0.87

Conclusion

The project's exploration of various NLP techniques for email spam classification highlights the importance of feature engineering in text classification tasks. Combining traditional text representation methods like BoW and TF-IDF with linguistic features derived from sentiment analysis, negation handling, and POS tagging can significantly enhance model performance. Specifically, models that included sentiment and POS tagging features generally outperformed those that did not, underscoring the value of incorporating semantic and syntactic context into classification models.

Negation handling, while conceptually valuable, did not consistently improve model outcomes, indicating that its effectiveness may depend on the specific characteristics of the text data and the interplay with other features. Moreover, the findings suggest that while TF-IDF is a powerful tool for highlighting important words, in some contexts, simple frequency counts (as used in BoW models) can be more effective for capturing relevant patterns in spam detection.

Overall, this project demonstrates the nuanced and multifaceted nature of NLP in practical applications like spam detection. The results emphasize the need for careful feature selection and combination to build robust and accurate classification models. Future work could explore more sophisticated NLP techniques, such as word embeddings and deep learning models, to further enhance the detection of spam emails.