**NLP Group Project Analysis and Insights**

**1) Data Exploration**

**Dataset Size:** 1956 entries with 2 columns (comment, label).

**Class Distribution:**

* Non-Spam (label = 0): 951 comments.
* Spam (label = 1): 1005 comments.
* The dataset is balanced between the two classes, this is beneficial for training a model without introducing significant bias.

**Null Values:** None. The dataset is clean and does not require handling missing values.

**2) Preprocessing**

**Text Preprocessing:**

* Converted text to lowercase.
* Removed punctuation.
* Tokenized the text into words.
* Removed stopwords (common words like "and," "the," etc.) using NLTK's English stopwords.

**Processed Text Examples:**  
Raw comments were converted to a simplified, processed format. For example:

* **Original:** "I really love this video! Thanks for sharing!"
* **Processed:** "really love video thanks sharing"

**3) Feature Extraction**

**Bag of Words (CountVectorizer):**

* Extracted 4180 unique features (words) from the dataset. Each feature represents a word found in the processed text.
* The feature matrix has a shape of (1956, 4180). 1956 Lines of comments and 4180 columns of uniquely identified words

**TF-IDF Transformation:**

* Transformed the Bag of Words representation to a TF-IDF (Term Frequency-Inverse Document Frequency) matrix, which scales the features based on their importance in the dataset.
* TF-IDF Matrix Shape: (1956, 4180).

**4) Model Training and Validation**

* **Classifier:** Multinomial Naive Bayes.
* **Cross-Validation:**
  + Used 5-fold cross-validation on the training set.
  + **Cross-Validation Accuracy Scores:** [0.8605, 0.8980, 0.8874, 0.8123, 0.8669]
  + **Mean CV Accuracy:** 0.8650 (86.5%).

This high cross-validation accuracy demonstrates that the model generalizes well during training.

**5) Model Testing**

* **Test Set Results:**
  + **Accuracy:** 87.73%
* **Classification Report:**
  + **Precision (Non-Spam):** 89%  
    The model is good at avoiding false positives when predicting Non-Spam.
  + **Recall (Non-Spam):** 88%  
    Most Non-Spam comments are correctly identified.
  + **Precision (Spam):** 87%  
    Spam detection is also strong, with minimal false positives.
  + **Recall (Spam):** 87%  
    Most Spam comments are correctly detected.
  + **Overall Accuracy:** 88%.

**6) Testing on New Comments**

**Analysis and Insights of the Classifier’s Predictions**

***"This is such a wonderful video! Thank you for sharing your insights. Truly inspiring"***

**Prediction**: Spam

**Analysis**:

The classifier incorrectly labeled this **non-spam comment** as **spam**.

**Possible Reasons:**

* Words like "wonderful", "sharing", and "inspiring" are likely associated with spammy comments in the training dataset, such as "Check out my channel for wonderful content" or "Sharing amazing offers".
* Polite spam comments in the dataset might have biased the classifier to mislabel genuine gratitude as spam.

I**nsight:**

* The model struggles with contextual understanding and over-relies on specific keywords without differentiating between genuine and promotional usage.

***“Katy Perry performance here is fantastic. Love the creativity in her work!”***

***“I can’t believe how much effort was put into this video. Dance was awesome!”***

***“This helped me understand the topic so much better. Your songs are amazing!”***

**Prediction**: Non-Spam

**Analysis**:

The classifier **correctly** labeled this as **non-spam**.

**Possible Reasons:**

* Words like “fantastic” and “creativity” are associated with positive fan feedback rather than promotional content.
* The training data likely contains many similar comments of fans praising performances or music.
* The comment has a clear focus on appreciating the effort and dance in the video, without any promotional or spam-like language.

**Insight**:

The model generalizes well for comments that are straightforward and do not include ambiguous

keywords like “sharing” or “subscribe”

***“Win a free iPhone by clicking the link below! Don’t miss this amazing offer”***

***Subscribe to my channel for exclusive content and free downloads!”***

**Prediction: Spam**

**Analysis:**

The classifier **correctly** labeled this as **spam**.

**Possible Reasons:**

* Words like “free, clicking, link” and “offer” are strong indicators of spam.
* This type of comment is likely well-represented in the training dataset, making it easier for the classifier to identify.

**Insight:**

The model effectively detects classic spam patterns with strong promotional language and links.

**General Observations**

1. Strengths of the Classifier:

* It performs well on detecting classic spam patterns (e.g., promotional language, calls-to-action, and links).
* It generalizes well for clear non-spam comments that lack ambiguous keywords.

2. Weaknesses of the Classifier:

* It struggles with ambiguous non-spam comments that contain keywords often found in spam (e.g., “sharing, wonderful, amazing”).
* It lacks contextual understanding, leading to misclassification of polite or gratitude- based comments as spam.

**Insights and Recommendations**

**1.** Use bigrams and trigrams in feature extraction to capture context. For example:

“Sharing knowledge” vs. “sharing link”

**2.** Threshold Adjustment:

Adjust the decision threshold for classifying comments as spam to reduce false

positives for ambiguous non-spam comments.

For example:

proba = clf.predict\_proba(new\_comments\_vectorized)

custom\_threshold = 0.7 # Classify as spam if probability &gt; 0.7

predictions\_custom = (proba[:, 1] &gt; custom\_threshold).astype(int)