SENTIMENT ANALYSIS SYSTEM

GROUP H

Ronakkumar Chavda – c0925342
Navdeep Kaur – c0929191
Rahul Rawat – c0928597
Aniket Sehrawat – c0928583
Raj Prasad Shrestha – c0924879
Devi Unni – c0928346

1. Introduction

This report summarizes the development of a sentiment analysis system using Bag of Words (BoW) and TF-IDF techniques. The goal was to preprocess text data, apply machine learning models, and evaluate their performance to classify sentiment effectively.

2. Data Challenges and Preprocessing Decisions

Data Challenges

- **Missing Values**: The dataset had 20 missing entries in the 'text' column, which could lead to errors during analysis.
- **Duplicates**: The dataset contained 960 duplicate text entries, risking bias and overfitting.
- **Noise**: Non-alphabetic characters, uppercase letters, and stopwords introduced noise.

Preprocessing Steps

Step	Action	Justification
Handling Missing Values	Dropped rows with missing text or sentiment labels.	Ensures data quality and avoids errors during analysis.
Removing Duplicates	Identified duplicates but retained them for this analysis.	Prevents bias but may need removal in future iterations.
Text Cleaning	Removed non-alphabetic characters, extra spaces, and converted to lowercase.	Eliminates noise and standardizes text for analysis.
Stopword Removal	Removed common English stopwords (e.g., "the", "and").	Reduces dimensionality and focuses on meaningful words.
Lemmatization	Reduced words to base forms (e.g., "running" → "run").	Ensures consistency in word representation.

3. Model Comparison and Best-Performing Model Selection

Feature Engineering

- **Bag of Words (BoW)**: Represents text as word counts.
- **TF-IDF**: Weights words based on their importance in the document and corpus.

Models Evaluated

Three models were trained and evaluated:

1. **XGBoost**
2. **Support Vector Machine (SVM)**
3. **Random Forest**

Performance Metrics (BoW Features)

Model	Accuracy	Precision	Recall	F1 Score
XGBoost	1.0000	1.0000	1.0000	1.0000
SVM	1.0000	1.0000	1.0000	1.0000
Random Forest	1.0000	1.0000	1.0000	1.0000

Performance Metrics (TF-IDF Features)

Model	Accuracy	Precision	Recall	F1 Score
XGBoost	1.0000	1.0000	1.0000	1.0000
SVM	1.0000	1.0000	1.0000	1.0000
Random Forest	1.0000	1.0000	1.0000	1.0000

Unseen Text Predictions

- **"This is a great product! I love it."** → Predicted as positive (1) by most models.
- **"I am not satisfied with the service."** → Predicted as negative (0) by all models.

Best Model Selection

All models performed perfectly on the test set, but **XGBoost** is recommended due to its robustness and scalability.