**CASE STUDY 1: IDENTIFY THE SENTIMENT OF TWEETS**

**Prepared by:** URMISHIKHA DASH  
**Date:** 2/4/25

**1. Executive Summary**

This report documents the development of a sentiment analysis model for electronic product-related tweets. The objective is to classify tweets into positive, negative, or neutral sentiments using natural language processing (NLP) techniques. Key components include data preprocessing, feature extraction, model training using machine learning algorithms, and performance evaluation.

**Key Results:**

* **Preprocessing:** Tokenization, stopword removal, and lemmatization were applied.
* **Feature Extraction:** TF-IDF vectorization was used.
* **Models Used:** Logistic Regression, Naïve Bayes, and Random Forest.
* **Best Model Performance:** Achieved **85% accuracy** using Logistic Regression.

**2. Objectives**

* **Data Cleaning:** Preprocess tweets to remove noise.
* **Feature Extraction:** Convert text data into numerical representations.
* **Model Training:** Train classifiers to predict sentiment.
* **Evaluation:** Assess models using accuracy, precision, and recall.

**3. Methodology**

**3.1 Dataset Preparation**

* **Source:** Collected tweet data related to electronic products.
* **Cleaning Steps:** Removed punctuation, URLs, special characters, and stopwords.
* **Preprocessing:** Tokenization, lemmatization, and lowercasing were applied.

**Dataset Statistics:**

| **Sentiment** | **Count** |
| --- | --- |
| Positive | 2,500 |
| Negative | 2,300 |
| Neutral | 2,200 |

**3.2 Feature Extraction**

* **TF-IDF Vectorization:** Converted cleaned tweets into numerical vectors.
* **N-grams Used:** Unigrams and bigrams for improved context capture.

**3.3 Model Training & Evaluation**

| **Model** | **Accuracy** | **Precision** | **Recall** |
| --- | --- | --- | --- |
| Logistic Regression | 85% | 0.84 | 0.85 |
| Naïve Bayes | 82% | 0.81 | 0.82 |
| Random Forest | 79% | 0.78 | 0.79 |

* Logistic Regression outperformed other models with **85% accuracy**.
* **Confusion Matrix Analysis:** Showed a balanced classification across sentiments.

**4. Results**

| **Metric** | **Performance** |
| --- | --- |
| Best Model | Logistic Regression |
| Sentiment Accuracy | 85% |
| Feature Extraction | TF-IDF |

**5. Discussion**

**5.1 Strengths**

Effective preprocessing improved model accuracy.  
TF-IDF performed well for feature extraction.  
Logistic Regression provided high accuracy.

**5.2 Limitations**

Model struggles with sarcasm and complex sentiments.  
Dataset is imbalanced in some sentiment categories.

**5.3 Recommendations**

Use deep learning models (LSTMs or Transformers) for better accuracy.  
 Expand dataset with more diverse samples.  
 Apply sentiment augmentation techniques.

**6. Conclusion**

The sentiment analysis model successfully classified tweets with **85% accuracy**. Future improvements include using deep learning techniques and handling complex linguistic nuances for better predictions.

**Deliverables Submitted:**  
Jupyter Notebook (NLP CS1.ipynb)  
Processed Dataset

**Appendix A: Sample Data**

| **Tweet** | **Sentiment** |
| --- | --- |
| "The new phone is amazing! 😊" | Positive |
| "Worst product ever! Waste of money." | Negative |

**Appendix B: Code Snippets**

# Logistic Regression Model Training

clf = LogisticRegression()

clf.fit(X\_train\_tfidf, y\_train)

# Model Prediction

y\_pred = clf.predict(X\_test\_tfidf)

# Accuracy Score

accuracy\_score(y\_test, y\_pred)