INFORMATION TECHNOLOGY UNIVERSITY

Machine Learning

Project Proposal

A Sentiment Analysis System

Powered by Ensemble Learning

Requested By: Aitsam Atif BSCE22012

Umair ul hassan BSCE22032

1. Project Title

" A Sentiment Analysis System Powered by Ensemble Learning"

---

2. Objective

Develop a robust sentiment analysis system using ensemble learning to classify text into **positive, negative, or neutral** categories. The system aims to:

- Provide actionable insights into customer opinions and brand perception.
- Demonstrate improved accuracy and generalization through ensemble modeling compared to individual classifiers.

---

3. Methodology

3.1 Data Collection & Preprocessing

- **Data Sources**: Publicly available text datasets (e.g., social media posts and comments (i.e Facebook ,X,Instagram).
- **Text Cleaning**:
  - Convert all text to lowercase.
  - Remove punctuation and special characters.
  - Tokenize text into words.
  - Apply lemmatization to reduce words to their root forms.

3.2 Feature Extraction

- Convert cleaned text to numerical features using **TF-IDF (Term Frequency-Inverse Document Frequency)**.
- Limit features to the **top 5,000 terms** to reduce computational complexity and avoid overfitting.

3.3 Model Development

- **Base Models**:
  1. **Naive Bayes**: Fast and effective for text classification.
  2. **Logistic Regression**: Strong performance in multi-class scenarios.
  3. **Support Vector Machine (SVM)**: Handles high-dimensional data effectively.
- **Ensemble Learning**:
  - Combine predictions using a **Voting Classifier** (majority voting).
  - Benefits: Improved accuracy, reduced overfitting, and leveraging strengths of diverse models.

3.4 Deployment for Predictions

- Preprocess new text (cleaning, tokenization, lemmatization).
- Convert to TF-IDF features.
- Predict sentiment using the trained ensemble model.

---

4. Tools & Technologies

- **Programming Language**: Python
- **Libraries**:
  - NLP: NLTK, spaCy (for preprocessing).
  - Machine Learning: scikit-learn (TF-IDF, models, ensemble).
  - Visualization: Matplotlib/Seaborn (performance metrics).
- **Environment**: Jupyter Notebook, Google Colab/cloud platforms (for scalability).

---

5. Expected Outcomes

- A deployable sentiment analysis system with **higher accuracy** than standalone models.
- Detailed performance reports (accuracy ≥ 85% on validation data).
- Insights into model interpretability and feature importance.

---

6. Conclusion

This project will deliver a scalable sentiment analysis solution that combines the strengths of multiple machine learning models. By leveraging ensemble learning, the system will provide businesses with reliable insights into public sentiment, enabling data-driven decision-making.

Requested By: Aitsam Atif BSCE22012, Umair ul hassan BSCE22032
Date: 25/2/2025