**To:** Prof. Ethan Davis

**Course:** Machine Learning Frameworks - 01

**Project Proposal:** Twitter sentiment Analysis

**URL:** https://www.kaggle.com/datasets/kazanova/sentiment140

**Group Members**

Alexander Sanchez - 200645900

Rahul Maddenapalli - 200638089

Vishal Patel - 200592864

**Problem Statement**

In today's digital landscape, social media platforms have become the primary venues for public discourse, with millions of users expressing their thoughts, opinions, and emotions daily. Twitter alone generates over 500 million tweets per day, creating an unprecedented volume of textual data rich with human sentiment and behavioral insights.

However, manually analyzing this vast amount of unstructured text data is practically impossible. Organizations, businesses, and researchers struggle to efficiently extract meaningful sentiment insights from social media conversations, limiting their ability to:

- Understand public opinion about products, services, or topics

- Monitor brand reputation and customer satisfaction

- Identify emerging trends and market sentiments

- Make data-driven decisions based on public sentiment

The challenge lies in developing automated systems that can accurately interpret the nuanced emotional context within human language, accounting for sarcasm, context, and varied expression styles.

**Project Goal**

Develop and deploy an intelligent sentiment analysis system capable of automatically classifying text sentiments with high accuracy, enabling real-time analysis of social media content and supporting informed decision-making processes.

**Project Scope & Objectives**

This project will deliver a comprehensive sentiment analysis solution through the following key objectives:

**1.Data Foundation**

* Data Acquisition: Utilize the Sentiment140 dataset from Kaggle containing 1.6 million labeled tweets
* Data Preprocessing: Implement robust cleaning, normalization, and feature extraction pipelines
* Quality Assurance: Ensure data integrity and handle missing or corrupted entries

**2. Analytical Intelligence**

* Exploratory Data Analysis: Conduct comprehensive statistical analysis to identify patterns, trends, and data characteristics
* Feature Engineering: Extract and optimize relevant linguistic features for model training
* Sentiment Distribution Analysis: Understand the balance and distribution of sentiment classes

**3. Model Development & Optimization**

* Multi-Algorithm Approach: Implement and compare multiple machine learning algorithms (e.g., Naive Bayes, SVM, Random Forest, Neural Networks)
* Performance Optimization: Fine-tune hyperparameters and model architectures for optimal accuracy
* Cross-Validation: Ensure model reliability through rigorous validation techniques

**Evaluation & Deployment**

* Comprehensive Evaluation: Assess model performance using multiple metrics (accuracy, precision, recall, F1-score, ROC-AUC)
* Comparative Analysis: Provide detailed performance comparison across different algorithms
* Model Selection: Identify the best-performing model for production deployment

**Expected Deliverables**

* Trained sentiment classification models with performance benchmarks
* Interactive dashboard for real-time sentiment analysis
* Comprehensive documentation and technical report
* Scalable codebase ready for production implementation