SCHOOL OF COMPUTING
Faculty of Engineering

Project Proposal Form MCST1043
Sem: 2 Session: 2024/25

SECTION A: Project Information.

Program Name:	**Masters of Science (Data Science)**
Subject Name:	**Project 1 (MCST1043)**
Student Name:	Gao Jingkai
Metric Number:	MCS241032
Student Email & Phone:	gaojingkai@graduate.utm.my & +60167101780
Project Title:	Knowing our choices: unveiling true voting patterns through machine learning (ML) and natural language processing (NLP) in European Parliament
Supervisor 1:	
Supervisor 2 / Industry Advisor(if any):	

SECTION B: Project Proposal

Introduction:

Recently, with the rapid growth of social media (especially Twitter), users around the world are Posting a large amount of emotional text on a variety of platforms every day. These contents not only reflect the individual psychological state, but also reveal the face of each group as a whole. Therefore, analyzing text data on these platforms is of great value for understanding user sentiment, platform policy adjustment, and academic research. However, the traditional NLP technology has some limitations in the application and semantic understanding of complex contexts. In recent years, pre-trained language models based on the Transformer architecture, such as BERT, have become very effective at understanding textual data. Therefore, this study aims to take advantage of this technological advantage, process and analyze social media text data, and build a set of models that can accurately predict the emotional tendency (positive or negative) of tweets.

Problem Background:

With the popularity of social media such as Twitter, users express their current views, attitudes and emotions about certain events, people and objects in words all the time. These contents not only have rich emotional color, but also contain great research value. However, most social media texts are characterized by strong subjectivity, leaping thinking, vague meaning and short text, which brings great challenges to traditional natural language processing technologies. In recent years, the emergence of pre-trained language models based on the Transformer architecture, such as BERT, has marked a major breakthrough in NLP and opened up new avenues to address these challenges.

Problem Statement:

While Transformer-based pre-trained language models such as BERT have made major breakthroughs in NLP in recent years, there are still limitations when applying them directly to analyzing text data from social media, especially Twitter. Many users use informal, sarcastic, cryptic expressions to express their emotions. This greatly limits the model to accurately judge the user's emotions. Therefore, this research aims to build an improved prediction framework that integrates semantic embedding and classical classification algorithms in Transformer to improve the accuracy, robustness and interpretability of the model for identifying text emotions in the social media environment, and provide more robust and accurate support for sentiment analysis tasks.

Aim of the Project:

This project aims to build an emotion prediction framework that integrates pre-trained Transformer language model with traditional machine learning algorithms to accurately identify users' emotional tendencies in Twitter text data.

Objectives of the Project:

1. To use a pre-trained model such as BERT to extract hidden emotional features from Twitter text.
2. To compare classic machine learning models (e.g., SVM, random forest, logistic regression) and identify which performs best in classifying tweets as "positive" or "negative".
3. To apply SHAP value analysis to determine which features most influence the model's classification of tweet sentiment.

Scopes of the Project:

1. This project uses only one public Twitter dataset for the experiment, and the dataset is real and valid.
2. This project focuses on the binary task of user emotions (positive/negative).
3. This project involves only use to the classification of the technology in traditional machine learning algorithms, such as SVM, random forests, logistic regression, not using deep neural network architecture, and using only preliminary training Transformer model as a semantic feature extraction tool.

Expected Contribution of the Project:

1. By combining BERT and traditional machine learning classification methods, an effective Twitter sentiment analysis framework is obtained.
2. If the method is feasible, can provide a complete set of media platform of user text data analysis process. It can provide a technical foundation for researchers.

Project Requirements:

Software:	Python programming language, Jupyter notebook, Anaconda
Hardware:	CPU: at least an Intel i5 (or equivalent)
	RAM: At least 16 GB (32 GB recommended for large datasets)
	Storage: Minimum of 200GB of free space
	GPU: Use it for training complex ML models
Technology/Technique/ Methodology/Algorithm:	Transformer Semantic Embedding (BERT-base-uncased)
	Classification algorithm: SVM, Random Forest, Logistic Regression
	Model evaluation indicators: Accuracy, Precision, Recall, F1-score
	Model interpretability techniques: SHAP value analysis, feature importance ranking

Type of Project (Focusing on Data Science):

[ ] Data Preparation and Modeling
[√] Data Analysis and Visualization
[ ] Business Intelligence and Analytics
[√] Machine Learning and Prediction
[ ] Data Science Application in Business Domain

Status of Project:

[√] New
[ ] Continued

If continued, what is the previous title?

SECTION C: Declaration

I declare that this project is proposed by:

[√] Myself
[ ] Supervisor/Industry Advisor ( )

Student Name: Gao Jingkai

Signature **Date** April 9, 2025

SECTION D: Supervisor Acknowledgement

The Supervisor(s) shall complete this section.

I/We agree to become the supervisor(s) for this student under aforesaid proposed title.

Name of Supervisor 1: .....

Signature **Date**

Name of Supervisor 2 (if any): .....

Signature **Date**

SECTION E: Evaluation Panel Approval

The Evaluator(s) shall complete this section.

Result:

[ ] FULL APPROVAL [ ] CONDITIONAL APPROVAL (Major)*

[ ] CONDITIONAL APPROVAL (Minor) [ ] FAIL*

***** Student has to submit new proposal form considering the evaluators' comments.

Comments:

Name of Evaluator 1: .....

Signature ..... **Date** .....

Name of Evaluator 2: .....

Signature ..... **Date** .....