SCHOOL OF COMPUTING
Faculty of Engineering

Project Proposal Form MCSD 6215
Sem:……. Session:……

SECTION A: Project Information.

Program Name: **Masters of Science (Data Science)**

Subject Name: **Project 1 (MCSD 6215)**

Student Name: Tian Fengshou

Metric Number: MCS20231029

Student Email & Phone: tianfengshou@graduate.utm.my

Project Title: Acquisition of Social Media Data and Sentiment Analysis

Supervisor 1:

Supervisor 2 / Industry Advisor(if any):

SECTION B: Project Proposal

Introduction:

As internet technology advances and intelligent mobile devices become widespread, social media has grown into an essential component of daily life for people.

The twitter, Weibo (a specific short text in the chinese APP- weibo), comment, reply and other short text content posted by users on social media platforms not only reflect their immediate emotions and attitudes, but also contain rich social information. Emotion analysis, as an important branch of natural language processing, aims to identify and extract emotional tendencies in text, and has significant application value in areas such as enterprise brand monitoring, market trend analysis, and public opinion monitoring.

Problem Background:

(1) Effective data acquisition methods: Commercial websites such as social media have anti crawler mechanisms, so corresponding technologies are needed to effectively obtain the data we need

(2) Effective data processing and management: The current social media data structure is complex and may include images, emojis, hyperlinks, videos, voice, etc. How to effectively extract, transform, and manage data will be a challenge.

(3)Efficient and low-cost sentiment analysis method: Modern social media data has a huge daily production volume. Due to

the enormous amount of data and the importance of sentiment analysis in public opinion monitoring, media public relations, and other fields, it is also a challenge to conduct sentiment analysis efficiently and at the lowest possible cost.

Problem Statement:

Although social media short text sentiment analysis has broad application prospects, it faces the following core issues: difficulty in data preprocessing, lack of targeted sentiment dictionaries, low accuracy in sentiment recognition, insufficient multimodal sentiment analysis, and difficulty in capturing sentiment evolution trends. This study aims to propose effective sentiment analysis methods to address these issues.

Aim of the Project:

Build a crawler system to selectively obtain data from social media or new media websites such as Weibo, Zhihu, Xiaohongshu, etc. Train a model using existing data, and then perform sentiment analysis using natural language processing algorithms.

Objectives of the Project:

Crawler system: at least two-thirds of Weibo, Zhihu, Xiaohongshu or other new media websites, as they plan to build cross platform tracking.

Data processing program: Multiple data processing programs suitable for reading, processing, converting, and storing data from different platforms.

Multiple sentiment analysis models: One sentiment analysis model that can be used for common Chinese social media network short texts.

Scopes of the Project:

This project just include short texts, does not include sentiment analysis of long texts or non social media texts.

The project does not involve sentiment analysis of voice or video data.

Excluding the work of developing mobile applications or user interfaces.

The results of the project will only be validated on specific social media platforms and cannot guarantee validity on other platforms.

Expected Contribution of the Project:

Build a crawler system that can crawl Weibo and Xiaohongshu, and crawl relevant data based on themes.

Develop an efficient data preprocessing tool that can automatically process short text data for specific platforms.

Develop an sentiment analysis model that achieves predetermined accuracy metrics on the test set.

Provide an emotional trend report showcasing key emotional changes monitored during the project period.

Project Requirements:

Software: Scrapy, Requests,Paddle, PyTorch, PyEcharts, Pandas

Hardware: Windows or Ubuntu, capable of running Paddle and Pytorch hardware effectively

Technology/Technique/ Methodology/Algorithm: Web crawler, NLP, Data visualization

Type of Project (Focusing on Data Science):

[Y] 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:

[Y] New

[ ] Continued

If continued, what is the previous title?