**STAT8037 Text Mining Group Project**

**Title and Headline Sentiment Prediction**

|  |  |  |
| --- | --- | --- |
| Luk Kin Fai | MDASC | 3035587124 |
| Ng Man Kit | MDASC | 2010069389 |
| Wong Chun Kit | MDASC | 2010011215 |
| Chan Cheuk Kai | MSTAT | 3035071826 |

**1. Project Objective**

Social media data are ubiquitous nowadays. People tend to share their emotion and opinions online and this provides golden opportunities for data professional to understand public opinion over certain topics. In this project, we will leverage on a kaggle dataset to explore stock sentiment through analyzing the headlines and titles. A text mining technique called “BERT” will be used to predict scoring

After building the models, we will evaluate which model is the best fit for prediction, and discuss future plans on stock trading.

Requirement of project:

• Objectives of the project (presenting the background of the project, the problem of the study, and project objectives).

• Data Sources: Description of text data and text preprocessing (including the source of data, the description of major features/variables, the quality of the data, and appropriate data preparation).

• Preliminary Findings (including the preliminary results of your analysis)

• Conclusions and Future Plans (describing the problem to be encountered and how they might be solved, etc.)

• References (such as research articles, books, book chapters, websites, etc.).

• Appendix (excluded from the 8 pages)

**2. Data Description and Preprocessing**

There are three available files to use, namely sample\_submission, train and test file. We will first focus on the train file and test the accuracy through test file. The data structure is shown as follows.

|  |  |
| --- | --- |
| **Feature** | **Description** |
| IDLink | Unique id for each post. |
| Title | The title of the post. |
| Headline | The headline of the post. |
| Source | The source from where it Originated. |
| Topic | The topic to which the post belonged to. |
| PublishDate | The date of Publishing. |
| Facebook | Weather it was present on Facebook or not. |
| GooglePlus | Weather it was present on GooglePlus or not. |
| LinkedIn | Weather it was present on LinkedIn or not. |
| **SentimentTitle** | **(Target) Title Score** |
| **SentimentHeadline** | **(Target) Headline Score** |

**2.1 Features data distribution:**

**3. Text Mining Method and Result**

**4. Conclusion**

**5. Discussion and limitation**

**6. Reference**