**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)

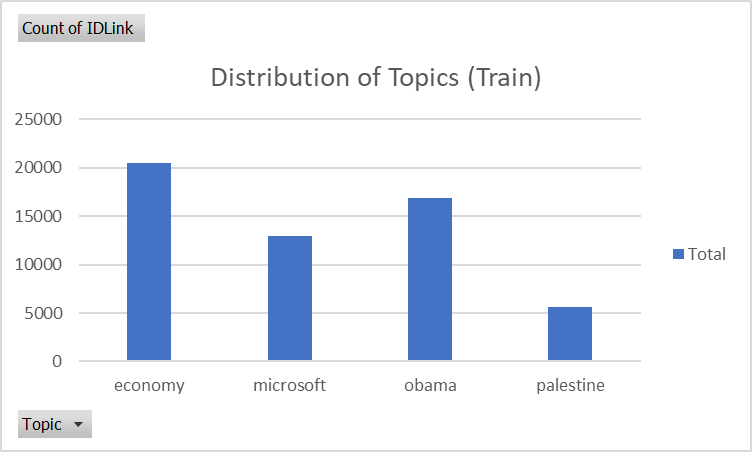
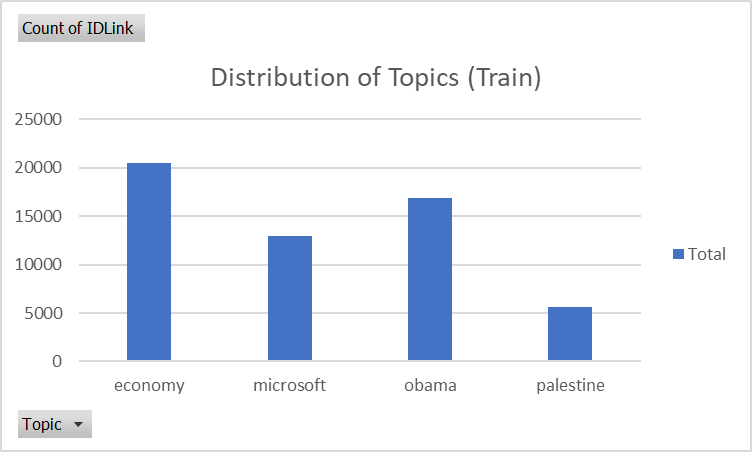
**2. Data Description and Preprocessing**

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 | # of views in Facebook (-1 = No posted) |
| GooglePlus | # of views in GooglePlus (-1 = No posted) |
| LinkedIn | # of views in LinkedIn (-1 = No posted) |
| **SentimentTitle** | **(Target) Title Score** |
| **SentimentHeadline** | **(Target) Headline Score** |

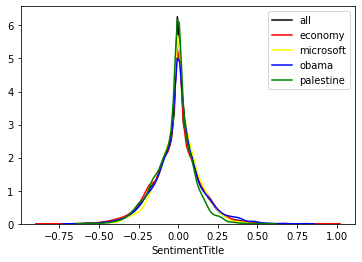
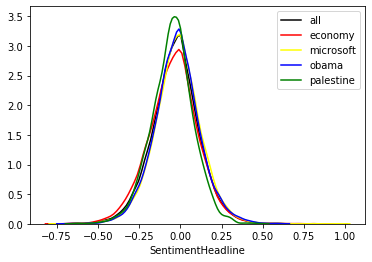
**2.1 Features data distribution:**

2.1a Distribution of Topic:



The data consist of 4 topics, including Economy, Microsoft, Obama and Palestine. Most of the Titles and Headlines are classified as Economy and Obama while few of them are classified as Palestine.

2.1b Distribution of Sentiment Title / Sentiment Headline:

The graphs show that there is not much difference on the distributions of Sentiment Headline among different Topic. All of them follow bell-shaped distribution which centered at zero. There is no skewness on the sentiment. However, some of the topic like Palestine have a higher density at zero, which indicate that more Titles / Headlines are classified as natural.

2.2 Data Preprocessing

Some data preprocessing techniques have been applied to clean the texts in Titles / Headlines.

2.2a Lower casing

Words in Upper case and Lower case are treated as a same word. For simplicity, all texts are converted to lower case.

2.2b Removing Punctuation

All punctuations have been removed from the texts as they are ambiguous.

2.2c Removing Stop words

The most common words like ‘a’, ‘is’, ‘the’ do not have useful information, and therefore are filtered out.

2.2d Removing frequent words / Rare words

Frequent words and Rare words may not have any predictive power in the model and therefore are removed from the text.

2.2e Correcting text

Spelling errors are corrected in the Headlines / Titles in order to make a better prediction.

2.2f Stemming / Lemmatization

Converting different form of words into root-stem by:

1. Removing suffices like ‘giving -> give’, ‘cats -> cat’ etc.
2. Reducing inflections or variant forms to base form, like ‘am, are, is -> be’

2.2c Text Distribution after Data preprocessing

**3. Text Mining Method and Result**

**4. Conclusion**

**5. Discussion and limitation**

**6. Reference**