Text Mining
Project Handout
2023

Market Behavior Prediction

Predicting stock market behavior from tweets

1. Project Objective

"Over time, major indexes go up and down based on internal and external factors. Performance like that excites investors, but typically in opposite ways. Constant gains lead some investors to expect more of the same. Others worry the good times are surely about to end. The former sentiment is sometimes called "bearish," while the latter is referred to as "bullish." Source: https://smartasset.com/financial-advisor/bullish-vs-bearish

The goal of this project is to develop an NLP model capable of predicting the Market sentiment based on tweets. In summary, with the NLP techniques you have learned during class, you must implement a classifier that receives tweets as inputs and is able to predict, for each tweet, if it describes a Bearish (0), Bullish (1), or Neutral (2) attitude.

The project should be developed using python 3 and libraries such as NLTK and Scikit Learn. Also, the project is simple and can be solved in various ways, which means there is no exact correct solution. Students should not use code from each other!

2. Group Rules

The project is to be developed individually or in groups of two (2) students.

3. Corpora

The data is divided in a file for training "train.csv", and another file for testing "test.csv":

- **Train** (9543 lines): Presents the tweets ("text") and the sentiment label ("label"). Each tweet can have 1 of the following labels: Bearish (0), Bullish (1), or Neutral (2)
- **Test** (299 lines): The structure of these dataset is the same as the train set, except that it does not contain the "label" column. You are expected to provide the predicted status (0, 1 or 2) for each tweet in this set and **we will compare your predictions with the actual (true) labels**.

4. Solution Requirements and Evaluation Criteria:

Your solution should present the following points:

1. **Data Exploration** (**0.75 points**): Here you should analyze the corpora and provide some conclusions and visual information (bar charts, word clouds, etc.) that contextualize the data.
2. **Corpus split** (**0.25 points**): You must apply some method to split your training corpus into train/validation sets to evaluate the performance of your model. You can also resort to K-Fold cross validation.
3. **Data Preprocessing** (**2 points**): You must correctly implement and experiment at least two (4) of the data preprocessing techniques shown in class (stop words, regular expressions, lemmatization, stemming, etc.).
4. **Feature Engineering** (**5 points**): You must correctly implement and experiment two (2) of the feature engineering techniques seen in class (BoW, TF-IDF, etc.).
5. **Classification Models** (**5 points**): You must correctly implement and test three (3) of the classification algorithms seen in class (KNN, LR, MLP, etc.).
6. **Evaluation** (**1 points**): You must evaluate your models resorting, at least, to Recall, Precision, Accuracy and F1-Score.

Moreover, the development of extra work (more techniques than the minimum required in the previous points and/or techniques not shown in class) is highly recommended and will account for a maximum of **4 points** divided as follows:

1. **Data Preprocessing – 0.5 points** for each extra method implemented and tested (maximum of 2 extra methods).
2. **Feature Engineering – 1.5 point** for using word embeddings.
3. **Classification Models – 1.5 point** for using LSMT or more advanced models.

5. Delivery Guide

In terms of the solutions developed (see **delivery template folder**), you must deliver:

1. One .pynb file (notebook), named NLP_XX (XX stands for the group number), containing the techniques you experimented following the structure mentioned in and your ready-to-run final solution.
2. A .csv file, named “Predictions_XX”, with the Ids of the test set and your predicted labels for the test set.

Additionally, you **must submit a PDF report** named “Report_XX”, documenting your work, with the following structure (other structures are also accepted):

1. **Data Exploration** – data presentation and explanation of the main finding from the exploratory analysis (accounts for **50%** of criteria **4.1**).
2. **Data Preprocessing** – explanation of the different preprocessing methods developed (accounts for **25%** of criteria **4.2** and **4.3**).

3. **Classification Models** – description of the models implemented (accounts for **25%** of criteria **4.5**)
4. **Evaluation and Results** – description of the performance of the models and main conclusions (accounts for **50%** of criteria **4.6**)

The PDF report should have a maximum of 5 pages describing the previous points. Exceeding this number will incur a **0.5-point penalty** for each extra page.

Any **extra work** developed must be clearly defined as such in the PDF report, or else it will not be considered for evaluation as extra work!

All files should be saved in a folder named "Group_XX". This folder (zip it if you need) must be submitted through Moodle's project submission section until **23h:59 of the 5th of April (Sunday)**.

Failure to deliver on time will incur a **1.0-point penalty** for each half-day late.
Failure to comply with the delivery guide will meet with a **0.5-point penalty**.

Final Notice:

We will compare your predictions with the actual Label from the test set ("test.csv").
The three (3) groups with the highest performance will receive points as follows:

- **1 point** for the group with the best model
- **0.5 points** for the group with the 2nd best model
- **0.25 points** for the group with the 3rd best model

Students may be randomly selected for an **oral defense** to access their knowledge.

Good luck with your project!