Mendel University in Brno

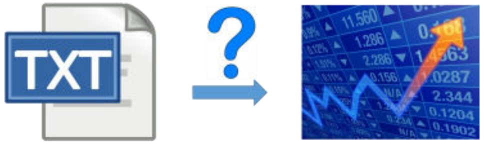
Acquiring and analyzing text data for financial markets

Written by Jonáš Petrovský, supervised by doc. František Dařena.

Faculty of Business and Economics

Goal

The thesis examines the connection between the content and sentiment of text documents published on the internet (which represent the public opinion) and stock price movements.



Used data

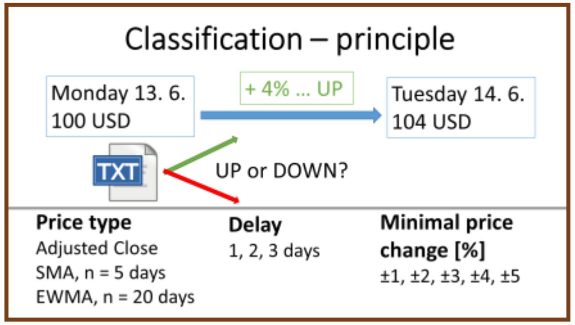
- Examined companies: S&P 500 + FTSEurofirst 300 = 800 companies.
- Data sources: More than 6 mil. documents acquired during a period of 8 months.
  - a) Yahoo Finance (news articles): 82 000.
  - b) Facebook (posts and comments on company pages): 135 000; 2.2 mil.
  - c) Twitter (statuses about companies): 4 mil.

Text classification

- This experiment examined, using supervised machine learning, the accuracy of price movement predictions based on a document’s text. Classification class represented the direction of the price movement of the related stock between two moments. Stock prices were smoothed using SMA (simple moving average) and EWMA (exponentially weighted moving average).
- Assigning class to a document was based on the relative price movement (see the image “Classification – principle” on the right) and on the determined value (v) of minimal price change. If the movement was in the constant interval (–v, +v), the document was discarded.
- Different values of three investigated parameters (price type, delay, minimal price change) generated 45 combinations (text files with documents of companies whose stock price complied with the parameter values). The documents were represented by three different weighting schemes and 6 classifiers were applied to the data. This means 810 experiments were conducted for each document type.
- The chart on the right shows, for each document type, an average accuracy of combinations with at least 500 documents. For each combination, the highest accuracy achieved by any of the 18 corresponding experiments was chosen. The combinations were ordered by accuracy from the highest and divided into groups. Column “top 10%” shows the average accuracy for top 10% combinations.
- The highest accuracy was achieved with EWMA price type, 1 day delay and 5% border.
- On average the best algorithm was LinearSVC. Each doctype had its best algorithm; decision tree CART was always the worst. All weighting schemes provided very similar results (±1%).

Sentiment analysis

- The goal of this experiment was to find out if the number of positive, negative and neutral news is connected with the growth, decrease or stagnation of stock price. For determining text sentiment VADER algorithm (together with newly created dictionaries) was used.
- The experiment was performed only for Adjusted Close price type and 2% border.
- The best accuracy (62%) was achieved for Yahoo articles and delay of 1 day.

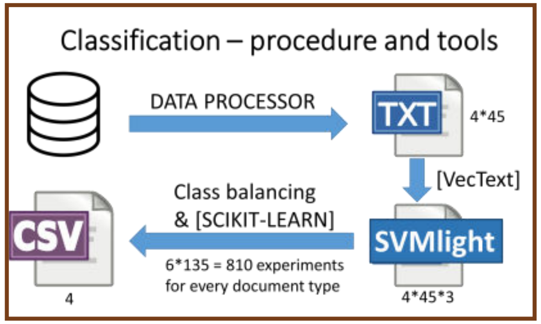


Classification – principle

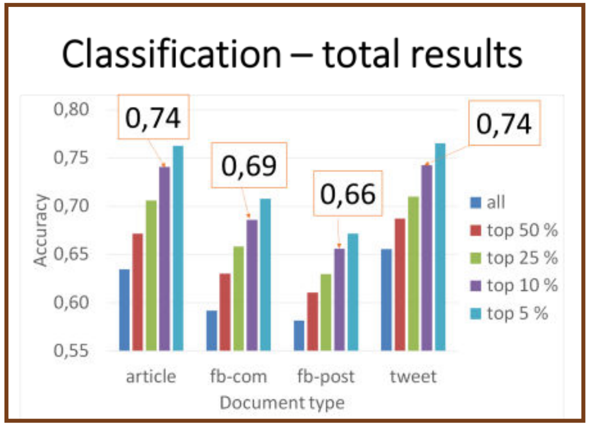
Price type
Adjusted Close
SMA, n = 5 days
EWMA, n = 20 days

Delay
1, 2, 3 days

Minimal price change [%]
±1, ±2, ±3, ±4, ±5



Classification – procedure and tools



Classification – total results

Conclusion

- The results of the classification show that if the price movement is (compared to the current trend) sufficiently substantial, there is a rather clear connection between the document’s content and stock price movement.
- This was showed by an accuracy of 70–80% for some configurations of the experiments.

Practical use

- Feature selection: Enables selecting the most important words for classification. This was used to create a new sentiment dictionary and to show which words indicate which movement.
- Classification: Companies can find out what news are connected with a specified price movement. Banks can analyze investor sentiment and explain price movements on the whole market.
- Sentiment analysis: It enables companies to evaluate how they are perceived by the public. Slightly edited algorithm and newly created sentiment dictionaries improve the results of this task.