MA5851 A3 Assessment Report Part One

Student: **Sacha Schwab**

Location: Zurich, Switzerland

Date: 3 December 2021

Github link to repository: https://github.com/sachaschwab/NLP-Clustering (https://github.com/sachaschwab/NLP-Clustering)

Background

Predicting how the stock market will move is a challenging issue due to many variables influencing an asset's price, such as interest rates, politics, and economic growth that make the stock market volatile and difficult to predict accurately. Technical analysis such as analysing company financial reports has its limitations and the need to enrich it with mining and interpreting unstructured text has become evident (Wigglesworth, 2017; Alzazah and Chen, 2020). Thus, the popularity of leveraging text mining and NLP techniques for market return analysis and predictions has significantly grown in recent years, from being used only by sophisticated quantitative hedge funds to a high demand in the broader market (Wigglesworth, 2017).

Most of the research and practice appear to focus on short-term effects of news articles on market prices, mostly by leveraging sentiment analysis. However, news stories can also be seen as events or stories that develop over time and that gain attention by market participants. There is a growing interest in clustering news into such events for further use by e.g. quantitative analysts (see e.g. Parse.ly's platform).

This project aims at providing base model allowing analysts to investigate the effect of event-grouped news articles, in particular on the price of cryptocurrencies. This is achieved by scraping an online news source, using NLP techniques to extract keywords and sentiments, and clustering the news articles based on these features.

Approach and architecture

To gather the relevant data, two web crawlers are employed: Firstly, the news articles crawler: News reports on cryptocurrencies and companies active in this space are collected from Yahoo Finance pages. Articles relating to Cryptocurrencies were collected over a period of 3 weeks using a web crawler leveraging the Beautifulsoup framework, which renders scaling and automatisation (to be used by e.g. job schedulers) easy. The workflow is such that the crawler detects the urls of the individual articles and then requests the html source of each article page. The html content is then parsed into title, article body and date metadata. The text data thus obtained was found appropriate for further handling i.e. for utilising NLP and machine learning by saving them as csv file.
Secondly, using Selenium, the list of all cryptocurrencies is crawled from another Yahoo page featuring tables with names of all cryptocurrencies.

In the second stage of the project, the text data were cleaned from symbols and stop words, tokenized and embedded into a representation of the importance of relevant words relative to the obtained text corpus. This was achieved by applying the TF-IDF algorithm. From this, keywords could be identified and noted along their importance. Further, using the Spacy framework and the crawled list of cryptocurrneices the text was scanned for persons, companies, locations and cryptocurrencies and added to the features. Finally for NLP, the sentiment of each article was derived using the Natural Language Tooklik (NLTK).

The purpose of the NLP tasks is to extract information appropriate as input for recognising and clustering the underlying events.

The numerical feature data added by the mentioned NLP processes were then used for clustering the articles into connected events. Following indications in the literature various clustering techniques were evaluated. Based on these evaluations, experiments and resource considerations, HDBSCAN was selected for the purpose of event clustering, however in this first iteration with limited success.

References

Alzazah and Cheng, 2020: Faten Subhi Alzazah and Xiaochun Cheng, “Recent Advances in Stock Market Prediction Using Text Mining: A Survey”, Intechopen 92253, June 2020

Bujari et al., 2017: Bujari A, Furini M, Laina N. On using cashtags to predict companies stock trends. In: 2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE; 2017. pp. 25-28

Cao et al., 2012: Tru H. Cao, et al., "Text clustering with Named Entities : A Model, Experimentation and Realization”, in: Data Mining: Foundations and Intelligent Paradigms, p 267-287, Springer-Verlag, Berlin Heidelberg, 2012

Capdeville, 2016: Joan Capdeville, et al., “Scaling DBSCAN-like Algorithms for Event Detection Systems in Twitter”, Department of Computer Architecture, Polytechnical University of Catalonia (UPC), December 2016

Ding et al., 2015: Xiao Ding, et al., “Deep Learning for Event-Driven Stock Prediction”, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), 2015

Ding et al., 2015: Ding X, Zhang Y, Liu T, Duan J. Deep learning for event-driven stock prediction. In: Twenty-Fourth International Joint Conference on Artificial Intelligence; 2015

Ellis, 2019: Daniel Ellis, “Using TF IDF to form descriptive chapter summaries via keyword extraction”, in: Towards Data Science online blog, November 2019

GDELT, 2015: The GDELT Project, GDELT 2.0: Our Global World in Realtime, February 2015, url: https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

Goyan et al., 2018: Archana Goyal, et al., “Recent Named Entity Recognition and Classification techniques: A systematic review” in: Computer Science Review 29 (2018) 21-

43, August 2018

Jabeen, 2018: Hafsa Jabeen, “Stemming and Lemmatization in Python”, in: Datacamp, October 2018

John and Vechtomova, 2017: Vineet John and Olga Vechtomova, “Sentiment Analysis on Financial News Headlines using Training Dataset Augmentation”, Association for Computational Linguistics, Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 869-873, Waterloo, UK, July 2017.

Krueger et al., 2019: Krueger, Samed, et al., “Event-Driven Strategies in Crypto Assets”, Conference Paper, April 2019

Li et al., 2021: Qian Li, et al., “A Comprehensive Survey on Schema-based Event Extraction with Deep Learning”, in: IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 9, august 2021

O’Gorman et al., 2021: Tim O’Gorman et al., “The Richer Event Description Corpus for Event-Event Relations” in: Computational Analysis of Storylines, Cambridge University Press, USA, 2021

Örs et al., 2020: Faik Kerem Ors, et al., “Event Clustering within News Articles”, Proceedings of AESPEN 2020, pages 63–68, Marseille, France, May 2020

Rusu et al., 2014: Delia Rusu, et al., “Unsupervised Techniques for Extracting and Clustering Complex Events in News”, in: Proceedings of the 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 26–34, Baltimore, Maryland, USA, June 2014

Velay and Daniel, 2018: Marc Veley and Fabrice Daniel, “Using NLP on news headlines to predict index trends”, Artificial Intelligence Department of Lusis, Paris, France, June 2018.

Wigglesworth, 2017: Robin Wigglesworth, “AI decodes trading signals hidden in jargon”, Financial Times online, New York, USA, October 2017

Xiang and Wang, 2019: Wei Xiang and Bang Wang, “A Survey of Event Extraction From Text”, School of Electronic Information and Communication, Huazhong, China, November 2019.

Xing et al., 2018: Xing, Frank Z., et al. “Natural Language Based Financial Forecasting: A Survey.” Artificial Intelligence Review, vol. 50, no. 1, June 2018, pp. 49–73!