Proposal for Master's thesis

Media Coverage on Financial Innovations: An NLP-Based Analysis

Author:	Simon Dolas
Student ID:	ds23m006
Study program:	Master Data Science
Advisor:	DI Dr. Elaheh Momeni-Ortner
Version:	0.2
Date:	08.09.2024

FH University of Applied Sciences TECHNIKUM WIEN

Problem area

The Oesterreichische Nationalbank (OeNB) shapes the economic development in Austria in the public interest. The main tasks of the OeNB include cash management, monetary policy, financial market stability, statistics, and payment systems. Additionally, it supports science and research, participates in the dissemination of economic and financial knowledge, promotes arts and culture, and reaffirms its commitment to diversity and sustainability. It is of particular interest to understand the public's stance towards certain topics served by the OeNB. The transparency platform for savings interest rates or the introduction of the digital euro are two examples where public perception is of great importance.

This master's thesis aims to conduct an analysis of public opinion on these financial topics using Natural Language Processing (NLP). To achieve this, articles from various media outlets will be scraped from the web and loaded into a pre-modeled database for analysis using sentiment analysis. The fact that media outlets generally aim to report neutrally poses a challenge in finding a suitable model. Additionally, user comments will also be examined. This also entails the challenge of identifying the stance and emotions of the users. For this purpose, a stance analysis (Mascarell *et al.*, 2021) and an emotion analysis (*mrm8488/t5-base-finetuned-emotion · Hugging Face*, 2023) will be conducted.

Data description

For the analysis of media coverage and user comments, various metadata and features are required in addition to the media articles. Therefore, as part of the thesis, the modeling and implementation of a suitable database for storing articles, user comments, and their associated metadata will be carried out. This will be populated using several web scrapers. Web scraping could pose a challenge as most media outlets try to technically prevent automated extraction or evaluation of their websites.

The Austrian Press Agency (APA), which offers a dashboard-product for media resonance analysis, manually and intellectually assigns sentiment to selected articles through its editorial team. The APA kindly granted permission to use these sentiments to validate the performance of the NLP model chosen in this masters thesis.

Research questions / hypotheses

To what extent can current Natural Language Processing (NLP) models contribute to the precise analysis of media coverage and public sentiment, as well as emotions regarding selected financial topics?

- What database model and DBMS are suitable for efficiently storing and managing data for this analysis?
- What strategies can be pursued to effectively address the technical challenges encountered by web scrapers?
- What NLP models are architecturally designed to
    1. evaluate media articles in terms of their tone, and what is their performance?
    2. determine the sentiment and emotions of users based on their comments?

Methods

To ensure that the database provides a clearly defined and consistent structure for storing and managing this extensive database, thorough database modeling is essential. After the requirements have been clearly defined, an appropriate DBMS is to be selected, and the database is to be implemented according to the modeling. The web scrapers are programmed in Python using the Selenium package, as it offers more flexibility in handling dynamic content. This is particularly useful since dynamic content can make automated access to the content more challenging. The ETL process, which populates the database with data using the web scrapers, starts daily automatically via cron jobs. To enable efficient filtering of the articles loaded into the database, additional features are generated using named-entity recognition or string matching.

Since BERT models often outperform neural networks in text classification tasks, such a transformer model from https://huggingface.co/ is used. (Guhr *et al.*, 2020)

The model oliverguhr/german-sentiment-bert could prove to be suitable for determining the tone of media articles, as it can also work with the German language. Additionally, for stance analysis or emotion analysis, an appropriate model needs to be selected. When the decision is made to pursue stance detection, the focus during model selection will be on zero-shot models. (Allaway and McKeown, 2020)

Once feature engineering is completed, the core questions can be answered through visualizations of the data.

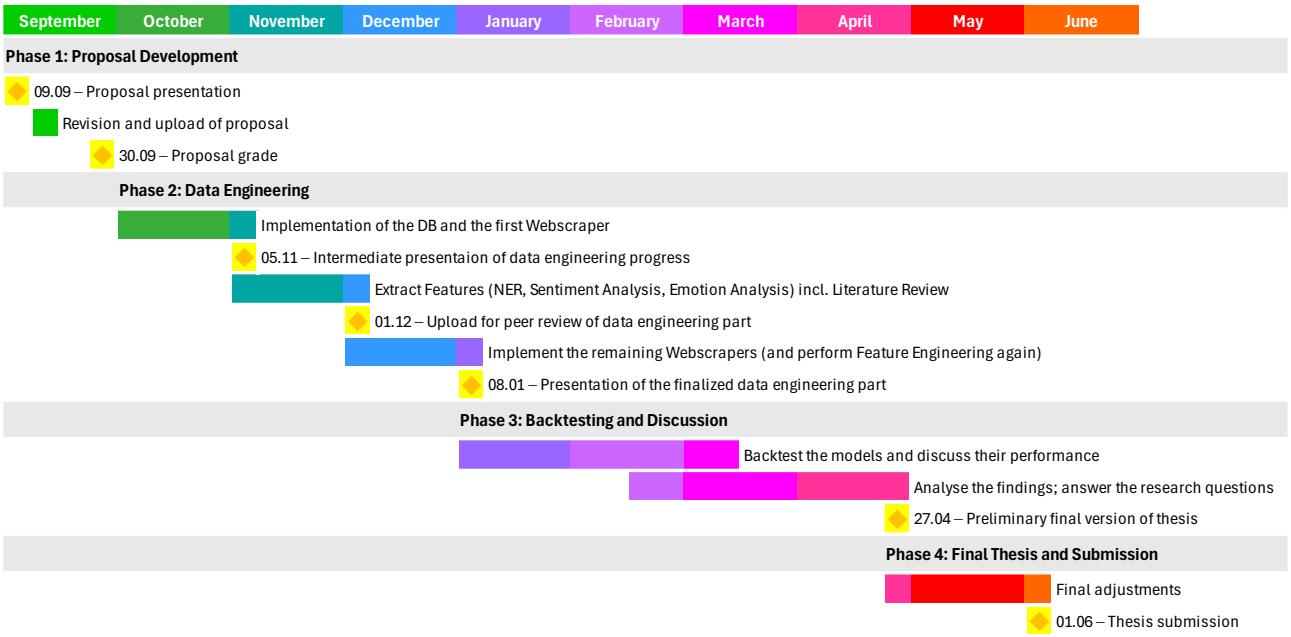
Expected results

The result of this work is a database that is populated through a daily scheduled ETL process using web scraping. Additionally, multiple classification models are employed for feature engineering. At least for the sentiment model, the performance can be determined and represented in the form of a confusion matrix, since the APA sentiments can be used as test data of the model.

Overview Research questions – methods – expected results

Research questions/ hypotheses	***Method(s) per research question***	***Expected kind of result per method***
Which database model and DBMS are suitable for efficiently storing and managing data for this analysis?	Literature Research	Implementation of a database solution
What strategies can be pursued to effectively navigate technical hurdles when using web scrapers?	Case Study	Webscrapers capable of handling dynamic content
Which NLP models are designed architecturally to evaluate media articles in terms of their tone, and what is their performance like?	Literature Research / Case Study	NLP model for sentiment analysis and its corresponding confusion matrix
Which NLP models are architecturally designed to determine the sentiment and emotions of users based on their comments?	Literature Research	List of suitable models for stance or emotion analysis

GANTT Chart


September
October
November
December
January
February
March
April
May
June
Phase 1: Proposal Development
09.09 – Proposal presentation
Revision and upload of proposal
30.09 – Proposal grade
Phase 2: Data Engineering
Implementation of the DB and the first Webscraper
05.11 – Intermediate presentaion of data engineering progress
Extract Features (NER, Sentiment Analysis, Emotion Analysis) incl. Literature Review
01.12 – Upload for peer review of data engineering part
Implement the remaining Webscrapers (and perform Feature Engineering again)
08.01 – Presentation of the finalized data engineering part
Phase 3: Backtesting and Discussion
Backtest the models and discuss their performance
Analyse the findings; answer the research questions
27.04 – Preliminary final version of thesis
Phase 4: Final Thesis and Submission
Final adjustments
01.06 – Thesis submission

Literature

Allaway, E. and McKeown, K. (2020) ‘Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations’. arXiv. Available at: https://doi.org/10.48550/arXiv.2010.03640.

Guhr, O. *et al.* (2020) ‘Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems’, in N. Calzolari et al. (eds) *Proceedings of the Twelfth Language Resources and Evaluation Conference*. *LREC 2020*, Marseille, France: European Language Resources Association, pp. 1627–1632. Available at: https://aclanthology.org/2020.lrec-1.202 (Accessed: 1 June 2024).

Mascarell, L. *et al.* (2021) ‘Stance Detection in German News Articles’, in R. Aly et al. (eds) *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*. *FEVER 2021*, Dominican Republic: Association for Computational Linguistics, pp. 66–77. Available at: https://doi.org/10.18653/v1/2021.fever-1.8.

mrm8488/t5-base-finetuned-emotion · Hugging Face (2023). Available at: https://huggingface.co/mrm8488/t5-base-finetuned-emotion (Accessed: 1 June 2024).