Automated Categorization of Form 8-K

About the Institute for Finance & Banking and the project

Research at the Institute for Finance & Banking aims at contributing to current issues in the fields of finance and banking. Our research activities are mainly quantitatively-empirically oriented. The focus is on

- the importance of banks for corporate finance (financial intermediation)
- the determining factors of the capital structure of companies
- ratings of companies
- corporate evaluation
- the determining factors of stock yields (asset pricing and cost of capital)

In academic teaching the focus is on theory-based empirical research in the areas of Finance and Banking as well as to provide quantitative methods.

Responsible for the project are Prof. Dr. Ralf Elsas and Moritz Scherrmann. Prof. Elsas is the chairholder of the Institute for Finance & Banking. Mr. Scherrmann is a 4th year PhD Student. Together they research the impact of ad hoc announcements of German listed companies on stock prices.

More precisely, they investigate the main drivers of stock market reactions and they look for ways to measure these drivers. Examples of these drivers are the surprise and severity of the announcement, its topic or its date (especially whether it is published inside or outside domestic trading hours), the issuers status and prestige or the country of the stock market (foreign or domestic).

Since ad hoc announcements are not standardized, it is challenging to gather meaningful measures for all these drivers from the raw text data. Hence, state-of-the-art textual analysis tools are required together with the manual creation of labeled datasets to end up with statistical models that are able to extract the needed information from the announcements.

Introduction:

Every public company in the U.S. is required to file a Current Report on Form 8-K after specific events. The types of information required to be disclosed on Form 8-K are generally considered to be “material”. That means that, in general, there is a substantial likelihood that a reasonable investor would consider the information important in making an investment decision.

Companies typically provide a number of 8-Ks throughout the year, whenever significant corporate events take place that trigger a disclosure. Companies must file 8-Ks promptly, rather than waiting until their next periodic report, such as the quarterly report or annual report. They are required to make most 8-K disclosures within four business days of the triggering event and in some cases even earlier. The public can find 8-Ks on the SEC’s EDGAR website (https://sec.report/Form/8-K).

The SEC has clearly defined when Form 8-K disclosure is required. For this purpose, there are a total of 9 superordinate sections, which are divided into different categories (see https://www.sec.gov/fast-answers/answersform8khtm.html). These sections can be seen as categories to which the disclosures belong.

Problem description:

We are interested in an algorithm which is able to assign suitable topics to disclosures automatically. Since multiple categories can be allocated to one disclosure, we are facing a so called *multi-label problem*. As already mentioned, companies filing a Current Report on Form 8-K are required to specify the section to which the disclosure belongs to, so one might assume that the category allocation is already done. However, the predefined categories of the SEC are partly a little broad. For example, section 7 contains topics like dividend announcements, Guidance, Earnings, etc. Section 8 contains all events that do not fit in the previous sections. These two sections alone account for 30% of all 8-K filings between 2015 and 2020. Therefore, the algorithm should be able to allocate categories to the disclosures that are independent of the section of the SEC and that are more fine-grained.

We already defined these categories and trained such an algorithm on a dataset consisting of the German counterpart of the 8-K filings: Ad-Hoc announcements. The algorithm is a BERT model followed by a classification layer. The task is to train a similar model for English data, i.e. 8-K filings.

Data:

Some of the German companies that are required to submit Ad-Hoc announcements do so using both languages, English and German. This gives us a parallel dataset, which contains numerous news pairs. A pair consists of an announcement with the identical content, but written in different languages. For the German part, we already have gold-labels available, which can be transferred to the English part. This setup allows the training of an English model.

We also provide a dataset consisting of 8-K filings. The performance of the English model has to assessed on that dataset. Of course, since we do not have any gold standard available here, a manual labeling has to be conducted too. However, since this dataset is only useful for testing purposes, the workload of this step is manageable.

The key steps of the assignment are:

- Compare the German ad-hoc announcements with the 8-k filings to get an intuition of the similarities and the differences
- Match English ad-hoc announcements to their German labeled counterpart
- Match sentences that are translations of each other. There exist already models that are able to do these kind of tasks, for example the “LASER”- model, the Multilingual Universal Sentence Encoder or Multilingual Sentence-BERT
- Train a BERT classifier which is able to assign one or more topics to each ad-hoc announcement
- Manually label a test set consisting of 8-K filings to test the model’s transferability on the new type of data. Analyze the result, find possible strengths and weaknesses of the method

As a possible final step, it will be interesting to link the 8-K filings to stock market reactions. Questions are:

- Which topics induce on average a significant positive or negative market reaction?
- Are there specific sentences or n-grams that imply material market reactions?

- Is there a systematic difference to Germany? (We provide all the results for the German case)
- Is there a difference in the information processing time between topics?

Additional Information:

- https://www.investor.gov/introduction-investing/general-resources/news-alerts/alerts-bulletins/investor-bulletins/how-read-8
- https://www.deltavalue.de/form-8-k-sec-filing/

Literature:

- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* (2019).
- Artetxe, Mikel, and Holger Schwenk. "Margin-based parallel corpus mining with multilingual sentence embeddings." *arXiv preprint arXiv:1811.01136* (2018).
- Yang, Yinfei, et al. "Multilingual universal sentence encoder for semantic retrieval." *arXiv preprint arXiv:1907.04307* (2019).
- Reimers, Nils, and Iryna Gurevych. "Making monolingual sentence embeddings multilingual using knowledge distillation." *arXiv preprint arXiv:2004.09813* (2020).
- Feuerriegel, Stefan, Antal Ratku, and Dirk Neumann. "Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation." *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2016.