DiFAn - Enhancing financial analysis with NLP

Timur Sheidaev, Danil Muzafarov, Nikita Pakunov, Alexandra Topalidi

May 2024

Abstract

This work aims at broadening the scope of stock analysis tools with the use of Natural Language Processing (NLP). By using sentiment analysis and Named Entity Recognition (NER), we try to bring more important details to the overview of the company, resulting in the analysis that is both easier to interpret and more thorough.

Project's code repository can be found here

1 Introduction

Financial sentiment analysis is a useful tool in the decision-making process of any investor, whether they know it or not. Some base their stock portfolios entirely on market sentiment, while others approach the topic more carefully, looking for companies that have been continuously overrated or underrated and combining their findings with more objective data. This work is an attempt to streamline the latter approach by integrating sentiment analysis into a service providing fundamental analysis based on Discontinued Cash Flow (DCF) and other financial models. We also try to enable a deeper analysis by providing connections between companies using entity recognition.

1.1 Team

Five members participated in the project - one advisor and four students:

Yuri Ichkitidze - Head of department of Finance at Saint-Petersburg HSE, CEO of difan.xyz and mentor for this project.

Timur Sheidaev - Data Engineer and Backend Developer at difan.xyz, collected the dataset and integrated the models into production pipelines. Worked on the sentiment analysis task.

Danil Muzafarov - Machine Learning Engineer, conducted experiments and exploratory data analysis, worked on the NER task.

Nikita Pakunov - Machine Learning Engineer, conducted experiments, worked on the NER task.

Alexandra Topalidi - Researcher, searched for articles and relevant work, proposed new use-cases.

2 Related Work

2.1 Sentiment analysis in finance

Sentiment analysis is the task of extracting sentiments or opinions of people from written language. Current approaches to solving this task can be roughly divided into two groups:

1) **Machine learning** methods with features extracted from text with "word counting";

2) **Deep learning** methods, where text is represented by a sequence of embeddings.

The former is a faster and cheaper approach, but it suffers from inability to represent the semantic information that results from a particular sequence of words, while the latter is often deemed as too "data-hungry" as it learns a much higher number of parameters [Marcus, 2018].

From a practical perspective, financial sentiment analysis differs from general sentiment analysis not only in domain, but also the purpose. The purpose behind financial sentiment analysis is usually guessing how the markets will react with the information presented in the text, which we try to leverage by providing larger quantities of data over a long time span, along with more traditional tools for financial analysis.

2.2 Named Entity Recognition (NER) in Financial News

Named Entity Recognition (NER) is a crucial component in our analytics framework, as it allows us to identify and classify stock tickers mentioned in financial news articles, providing deeper insights into the relationships and impacts of various companies and entities within the market.

In the end product, the result of our NER task is a list of tickers (companies), that are mentioned in the article and are present in the project's data warehouse (so that the user can do further research on them). There are several approaches to NER tasks, including:

1) **Rule-based approaches**: these rely on a set of predefined rules and patterns to identify entities. While these methods can be effective, they often lack generalizability and require significant manual effort to create and maintain the rules.

2) **Traditional Machine learning:** these methods involve training models on labeled datasets. Popular techniques include Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), but they often require a high level of expertise to implement correctly and tend to fall behind deep learning.

3) **Deep learning:** the advent of deep learning has significantly improved NER performance. Transformer models especially, like BERT and GPT, which leverage large-scale pretraining on diverse text corpora followed by fine-tuning on specific datasets, have achieved state-of-the-art results.

For our project, we utilized a BERT-based transformer model fine-tuned for NER tasks.

2.3 Prerequisite datasets and models

Perhaps the most relevant dataset regarding sentiment analysis in finance is the Financial PhraseBank [Malo et al., 2013], which contains around 5000 financial news titles labeled in 3 sentiment classes. This dataset is especially good because of a high number of annotations per record and good annotator agreement numbers (see Table 2.3 below)

Agreement level	Positive	Negative	Neutral	Count
100%	%25.2	%13.4	%61.4	2262
75% - 99%	%26.6	%9.8	%63.6	1191
66% - 74%	%36.7	%12.3	%50.9	765
50% - 65%	%31.1	%14.4	%54.5	627
All	%28.1	%12.4	%59.4	4845

Table 1: Distribution of sentiment labels and agreement levels in Financial PhraseBank [Malo et al., 2013]

In particular, this dataset was used by Prosus along with another large corpus of data - the Reuters TRC2 [1] - to train Finbert [Araci, 2019] - a fine-tuned BERT model, which is still one of the most popular models for the task.

3 Model Description

3.1 Sentiment Analysis

To accomplish the task, transformer architectures were chosen because of their great context awareness and ease of use with the Huggingface transformers library. BERT models are especially good at the task of sentiment analysis due to their bidirectional architecture.

The downside of being "data-hungry" due to a large number of parameters is mitigated by avoiding training the entire model from scratch and opting instead to fine-tune a pretrained model.

Other downside of deep learning methods is the cost of inference, which is also slightly higher. To lower the footprint, knowledge distillation can be used. During the process of distillation, a smaller "student" model is trained to mimic the behavior of a larger, more complex model, known as a teacher model. As a result, we get a lightweight model with most of the knowledge from the larger model.

Concluding the requirements and limitations, we have settled on a distilled version of RoBERTa-base[Liu et al., 2019], fine-tuned on the aforementioned FinancialPhraseBank and distilled down to 85M parameters to run on the CPU.

The sentiment analysis pipeline performs model inference on the article's title and content concatenated together, returning a softmax value for each

[1] This dataset can be acquired for academic purposes here: https://trec.nist.gov/data/reuters/reuters.html

class - negative, neutral and positive. One of the downsides of "just feed the entire text and get a score" approach is its naiveness - the model may not catch the whole context and/or misinterpret the targets of some adjectives.

3.2 NER - Mentioned Companies Extraction

For our NER tasks, we employed a BERT model in a SpanMarker framework that was fine-tuned on a multinerd dataset[Tedeschi and Navigli, 2022] - a diverse, 2.68M row dataset for fine-grained named entity recognition.

Since the articles are usually cased correctly, we have opted to use the cased version for better performance on limited resources. Being multilingual, this model may also come in handy when we scale our analysis beyond the western exchanges on, for example, the Russian market.

Our NER pipeline is a bit more complicated and involves the following steps:

1. **NER on text:** extract all entities from the text using the SpanMarker-BERT model

2. **Filter by company names:** from the extracted entities, filter those that correspond to company names. This step reduces noise and focuses the analysis on relevant financial entities.

3. **Fuzzy search:** perform a fuzzy search to match the extracted company names with the full company names in our database. This step accounts for variations and abbreviations in company names (e.g., "Boeing" -> "Boeing Co.").

4. **Retrieve ticker:** obtain the ticker symbol for the matched company name from our database. This involves querying our financial database to link the company name with its respective ticker symbol, ensuring accurate identification and tracking.

As a result, for each news article we get a list of tickers that are somehow associated with the main ticker about which the article is. Inevitably, the main ticker itself also gets into that list as an entity, but gets removed as a final step to avoid self-referencing.

4 Dataset

For this project, two API endpoints were connected - Financialmodelingprep and EOD Historical Data. These API's provide plenty of historical (since 2008) and real-time data while being relatively affordable[2].

Dataset was compiled in an SQL (MySQL flavor) database. For indexing, we have created an MD-5 hash column `article_id` encoding the article's title and content, which made running SQL queries on the dataset possible with reasonable speed. Table schema can be found in Tab. 2

After initial compilation from the API endpoints, the dataset was cleaned from articles that did not have any specific companies mentioned, e.g. "Top 5 Stocks That Gained This Morning", "Stocks That Hit 52-Week Highs On

[2] Both sources have a free tier subscription as well, suitable for research purposes

Name	**Type**	Comments
`id`	bigint	Autoincrement column for indexing
`article_id`	binary	MD5 hash of `title+content` for indexing
`date`	date	date of publication
`ticker`	text	Main stock in article
`publisher`	text	Article author
`url`	text	Link to source
`title`	text	
`content`	text	
`tickersMentioned`	text	Other companies mentioned in article
`neg`	float	Sentiment analysis (negative score), softmax
`neu`	float	Sentiment analysis (neutral score), softmax
`pos`	float	Sentiment analysis (positive score), softmax

Table 2: Dataset table schema

Wednesday" with little to no content, which made up almost 500000 records, or about 17% of the dataset. The easiest way to clean the dataset from such records turned out to be duplicate value counts:

```
SELECT article_id, COUNT(article_id) FROM news.dataset GROUP BY article_id;
```

returns `article_id` counts. After that, it is a straight-forward process of removing records with counts > 1, since meaningful articles rarely match exactly.

Furthermore, the dataset was cleaned from articles with short titles (usually having the stock ticker as a title), which removed another 1100 records, and articles with generic URLs that could not be traced to the original article (e.g. https://www.reuters.com/business/) by using a `GROUP BY` query similar to the duplicate `article_id` query, removing another 1400 records.

Overall the compiled dataset contains 2.4 million unique records for 17172 stock tickers and is growing at around 900 new articles each day. Out of all the companies initially loaded we are currently keeping track (performing inference) of only 1537 companies from major stock indices - SP500, NSDQ100 and HangSeng due to limitations in computational resources. Distribution of records over time can bee seen on Fig. 1.

Notably, we can see a large jump in the number of published articles during the COVID-19 pandemic. This correlates not only with the impact of said pandemic on global economy[Şenol and Zeren, 2020], but also with the increased interest of general public in the stock market[He et al., 2020].

One of the primary goals for future work on this dataset is to acquire data on previous years: having news on major economic events such as the 2008 housing crisis and the 1999-2000 dotcom bubble would provide valuable info for further experiments with sentiment analysis on a larger time span.

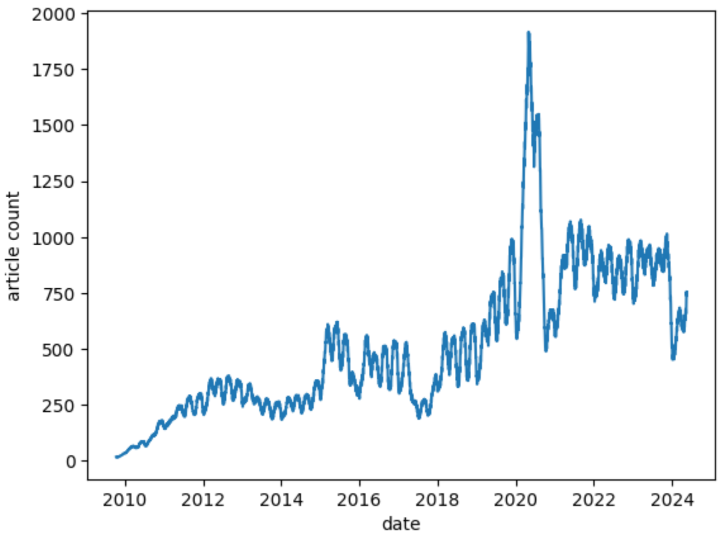


Figure 1: Article distribution (smoothed with 30-day moving average)

5 Experiments

While we did not train our own models for the described tasks, extensive research has been conducted to select a capable model (within our computational requirements of CPU-only inference), since quite a lot of work has been already done in this field.

In case of the NER task, it turned out that the nature of financial news helps greatly with the quality of predictions - usually the mentioned companies are stated clearly, sometimes already with tickers. Because of that, we opted to use a relatively small bert-base-sized model with a fine-tune on a broad dataset. On multinerd, the span-marker-bert-base-multilingual-cased-multinerd gets an F1-score of 0.952 for English language and 0.926 overall.

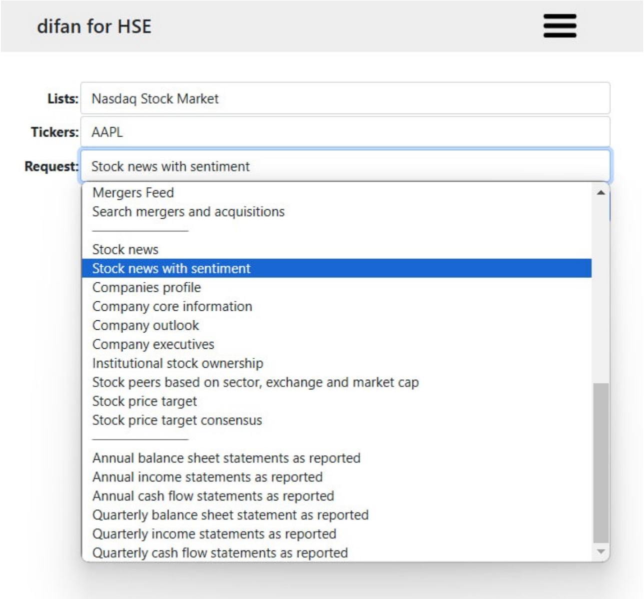
For the sentiment analysis task, we had to dig deeper: despite the aforementioned Finbert still being used widely, we searched for a smaller and better performing model. Tab. 3 below describes the models considered.

Name	**Training Dataset**	**Accuracy**	**F1 Score**
FinBERT	Financial PhraseBank	0.86	0.84
TOMFINSEN	Financial PhraseBank	0.87	0.87
DistilRoBERTa by nickmuchi	Financial PhraseBank + Financial Classification	0.88	0.89
DistilRoBERTa by mrm8488	Financial PhraseBank	0.98	0.97

Table 3: Comparison of different financial sentiment analysis models

6 Results

From the product point of view, the result of our work is a new data endpoint for the project's data terminal. This terminal serves two purposes: firstly, it replaces the now unavailable Reuters Eikon (Refinitiv Eikon) terminal for students of the Higher School of Economics (hence the "difan for HSE" title in the UI on Fig. 2) in Saint-Petersburg, Russia and some other institutions (and we are pleased to know that this dataset is already getting used in some student's theses); secondly, this application serves as a staging ground for the main project: here we have easy access to our experiments and can evaluate our hypotheses with new data.



Lists: Nasdaq Stock Market

Tickers: AAPL

Request: Stock news with sentiment

Mergers Feed
Search mergers and acquisitions
Stock news
Stock news with sentiment
Companies profile
Company core information
Company outlook
Company executives
Institutional stock ownership
Stock peers based on sector, exchange and market cap
Stock price target
Stock price target consensus
Annual balance sheet statements as reported
Annual income statements as reported
Annual cash flow statements as reported
Quarterly balance sheet statement as reported
Quarterly income statements as reported
Quarterly cash flow statements as reported

Figure 2: Difan data terminal user interface. "Stock news" and "Stock news with sentiment" are the endpoints developed in this work

In terms of the main project, the stock analytics tool, there remains some work to be done regarding integration of the new data into the stock analysis report. While the new data is useful, we don't want to overload the users with raw data in such form; that's why the next big task of this project will be to summarize the scores / relations of a company over the recent time period and write that in a form of key points. We have already done some experiments with simpler techniques (see project's repositiory), and it looks quite promising.

6.1 Example outputs

Below we provide a couple of processed records on Apple Inc. (AAPL) stock:

ticker	AAPL
publisher	investopedia.com
title	Stanley Druckenmiller Bet On Biotech, Financial Services in Q1, Trimmed Nvidia Stake
content	Billionaire Stanley Druckenmiller's Duquesne Family Office took on new positions in Apple Inc. (AAPL), Reddit Inc. (RDDT), and a range of biotech and financial services companies in the first quarter of the year, according to a recent 13-F filing.
url	link
tickersMentioned	["NVDA", "RDDT"]
sentiment (neg | neu | pos)	0.000676 | 0.999808 | 0.000124

Table 4: Output sample 1 - Neutral article

ticker	AAPL
publisher	InvestorPlace
title	Apple Stock Alert: Warning! AAPL's Problems Are Getting Worse.
content	Apple's (NSDQ: AAPL) first-quarter financial results contained some very discouraging signs, making the stock a sell in the near-term. Up until now, it has been easy to defend Apple stock.
url	link
tickersMentioned	None
sentiment (neg | neu | pos)	0.998459 | 0.000545 | 0.000996

Table 5: Output sample 2 - Negative article

ticker	AAPL
publisher	InvestorPlace
title	3 Warren Buffett Stocks That Look Irresistible
content	Warren Buffett's Berkshire Hathaway (NYSE: BRK.B) portfolio is rich with what the Oracle of Omaha would refer to as "wonderful" businesses. Though Berkshire's portfolio has skewed away from technology stocks in prior decades, one can't help but notice that the modern-day Berkshire portfolio has pivoted towards the modern age.
url	link
tickersMentioned	["BRK.B"]
sentiment (neg | neu | pos)	0.998459 | 0.000545 | 0.000996

Table 6: Output sample 3 - Positive article

7 Conclusion

In the course of this work, we have collected a large dataset of financial news and processed it with state-of-the-art deep learning models in order to extract valuable information, such as sentiment and relationships between companies through news. The results of this work are already used by students in their research, and will be further analysed to potentially integrate them into the financial analysis on the difan.xyz website.

References

[Araci, 2019] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models.

[He et al., 2020] He, Q., Liu, J., Wang, S., and Yu, J. (2020). The impact of covid-19 on stock markets. *Economic and Political Studies*, 8(3):275–288.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

[Malo et al., 2013] Malo, P., Sinha, A., Takala, P., Korhonen, P., and Wallenius, J. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts.

[Marcus, 2018] Marcus, G. (2018). Deep learning: A critical appraisal. *CoRR*, abs/1801.00631.

[Şenol and Zeren, 2020] Şenol, Z. and Zeren, F. (2020). Coronavirus (covid-19) and stock markets: The effects of the pandemic on the global economy. *Avrasya Sosyal ve Ekonomi Araştırmaları Dergisi*, 7(4):1–16.

[Tedeschi and Navigli, 2022] Tedeschi, S. and Navigli, R. (2022). MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.