NLP for analysing signals in news

Cooperation company project for RegHub UG

Neelesh Bhalla + Ananya Agarwal
(Frankfurt School of Finance & Management)

x

NLP for analysing signals in news

Management summary

The task

- News content processing, analysis, categorization and summarization.
- Annotation tool development.
- LLM inferencing with unlabeled news.
- Testing on a selected news sample.
- Results evaluation.
- Potential improvement strategies.

The dataset

- Financial news scraped from the web.
- Significantly enriched dataset; includes: (1) raw news data, (2) metadata crucial for in-depth analysis such as publication dates, topics, and RegHub-generated summaries.
- News items range from Jan'22 to Oct'23.
- 8.5k German and 6k English news items; we focus on English news.

The method

- LLMs' inherent capabilities of recognizing named entities, summarizing texts were exploited.
- Similarities between embeddings were used for supervised classification.
- Classification was also attempted by implementing topic modeling.
- Temporal signals were captured by combining a set of tasks enabled by classic NLP and modern LLM realm.

The findings

- Manual categorization is expensive and can be replaced by AI methods at lower cost and higher accuracy.
- Categorization accuracy is hard to measure, as several categories are often applicable to one news (at least for the dataset considered byRegHub).
- A lot of data cleaning can be required if the structure of news changes in anyway.
- Fine-tuning models is computationally expensive and skilled talent is rare in the labor market, but significant improvements are possible even compared to state-of-art LLM-models.
- Smaller, locally persisted LLMs can deliver phenomenal performance and eliminate the need of using API based inferencing on commercially available large LLMs.

NLP for analysing signals in news

Content

Overview

Task

- Background, project objective, scope and our approach

Dataset

- Data description and pre-processing
- Exploratory data analysis

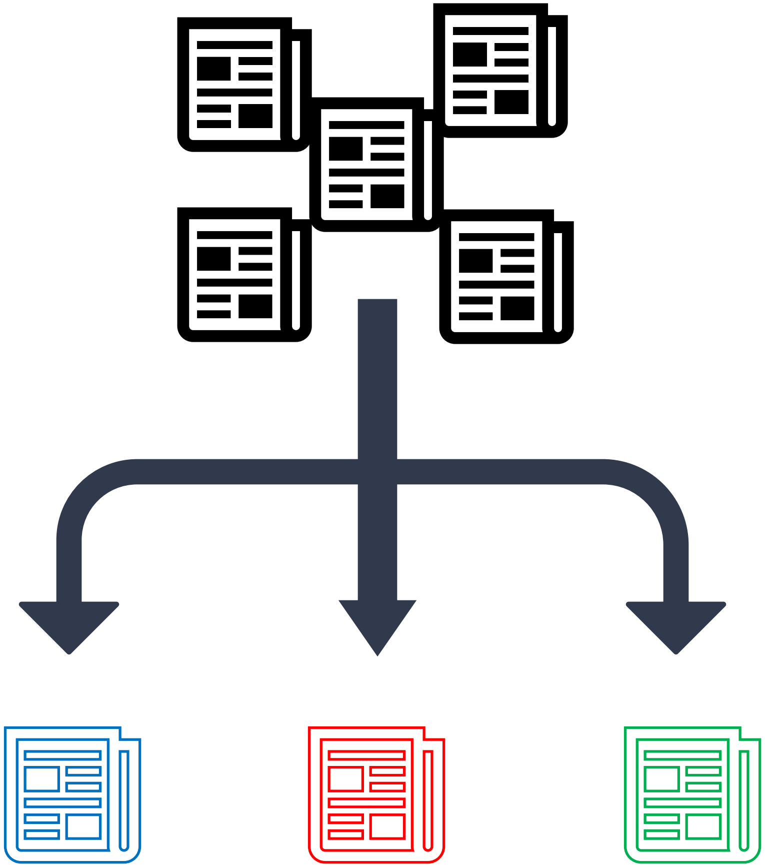
Method

- Model selection
- Extracting company names
- Spotting signals within news
- Ranking news items
- Unsupervised topic modelling
- Capturing temporal progress in news signals

Results

- Cost-compute-latency estimate analysis
- Challenges and learnings
- Summary and future outlook

Task



1. Task

Background, project objective, scope and our approach

Background, project objective and scope

- RegHub UG collects business news from the web. These news discuss (among others) financial, technological, strategic and regulatory topics around companies of strategic importance to RegHub's clients.

- From within this collected raw news texts, RegHub aims to provide the most accurate and concise business intel pertinent to companies of interest.

- ***Project objective***: *"To present a proof of concept around extracting meaningful data points (or business 'signals') from a set of news relevant to RegHub's clients, within a certain timeframe."*

- ***What is not working?***: (1) less efficient, 'keywords-based' supervised news *classification;* (2) hallucinated news *summarization* by Open AI GPT-4 apis; (3) missing standard workflow for *temporal analysis of news based textual data*.

Our approach

- We worked on the PoC using a sample dataset provided by RegHub. The project extensively exploited classic and modern NLP methods to fulfill the project objective.

- To name a few, tricks and techniques like '***comparing news similarities using sentence embeddings***', '***feeding in-context prompts to locally persisted small LLM***' and '***topic modelling using open-source libraries***' were carefully brought into use.

- Following our experimental research, the groundwork for an internal implementation at RegHub is laid.

Dataset

2. Dataset

Data description and pre-processing

News dataset description

- Financial news scraped from the web provided by RegHub.
- Significantly enriched dataset; includes: (a) raw news text and (b) 27 other columns of metadata crucial for in-depth analysis such as publication dates, topics, and RegHub-generated summaries.
- News items range from Jan'22 to Oct'23.
- 8.5k German and 6k English news items; we focus on English news.

Data preprocessing

- For accurate language filtering in our dataset, we utilized the ***'langdetect'*** library to analyze the news titles, as the language column alone didn't consistently represent the actual content language. Preprocessing revealed that out of the initial 6040 news content entries, 3868 were identified as English.
- In our topic modeling pre-processing, we applied stop words removal to enhance the model's effectiveness. Eliminating common words like 'and', 'the,' etc., aids in focusing on meaningful terms, improving the interpretability and relevance of topics extracted from the dataset.
- Owing to the advanced context-aware capabilities of Llama2, stop word removal is omitted in our data pre-processing, allowing the model to grasp contextual nuances. Additionally, the average word count in our news corpus without pre-processing comfortably aligns with Llama2's permissible context size, eliminating the need for further adjustments.

2. Dataset

Exploratory data analysis (1/3)

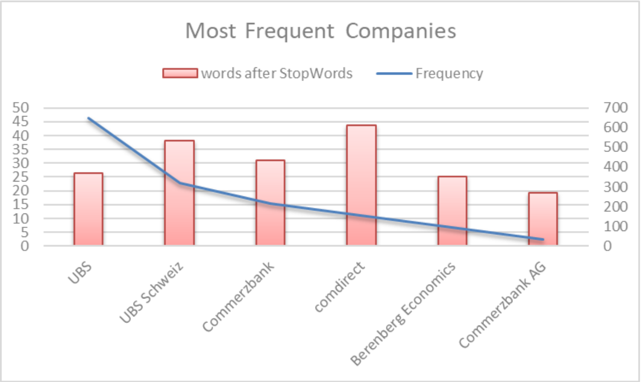
News Quantity vs. Content Quality

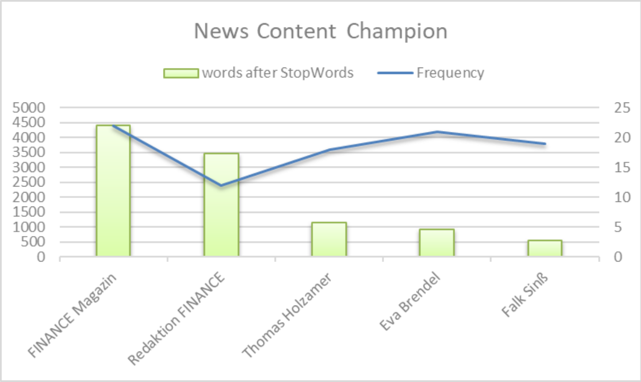
- Analysis reveals that 75% of the dataset focuses on a small set of key companies like UBS, Commerzbank, Comdirect, and Berenberg.

- Surprisingly, the average content length for these major companies is relatively short, averaging just 25 words after removing stopwords.

- Reporters like Finance magazin and Eva Brendel stand out with word counts exceeding a thousand, reflecting rich and in-depth content contributions.

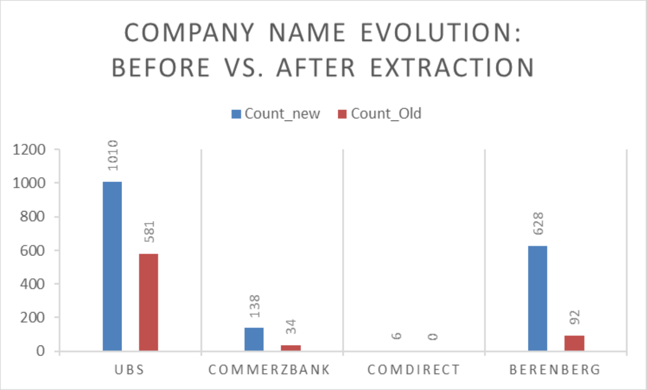
- Considering the vast difference, careful selection of authors is advised. Reporters like Finance Magazine, with an average length of 4397 words after removing stopwords, can provide more substantial content for in-depth analysis.

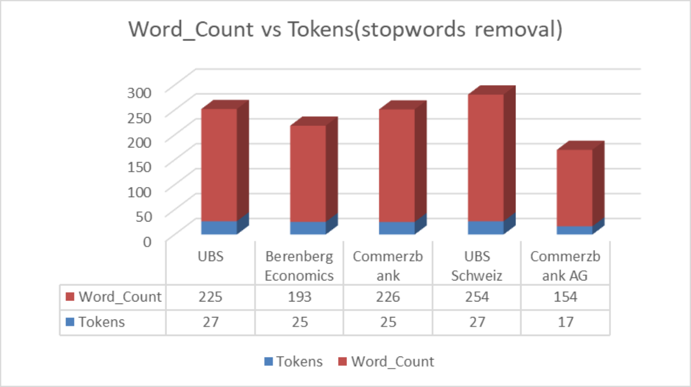




2. Dataset

Exploratory data analysis (2/3)





Company Name Evolution

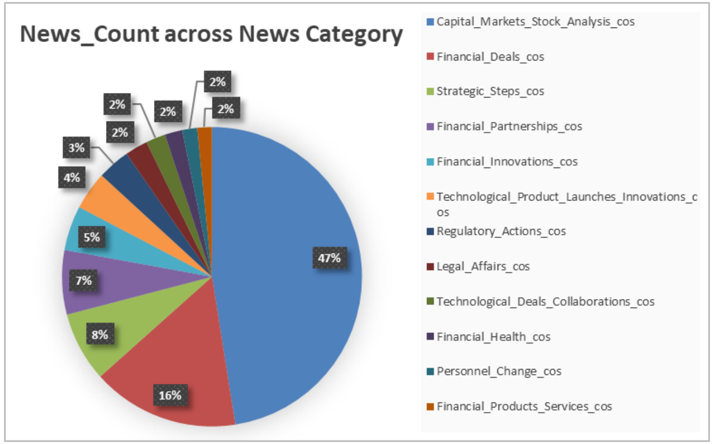
- For English news content dataset, the new data extraction process meticulously captures and populates company names, resulting in a substantial reduction in empty columns and providing a richer dataset for analysis.

Word Count vs Tokens

- A comparative analysis of word count with and without stopwords removal - unveiling that a substantial 90% of our word count resides in stopwords.

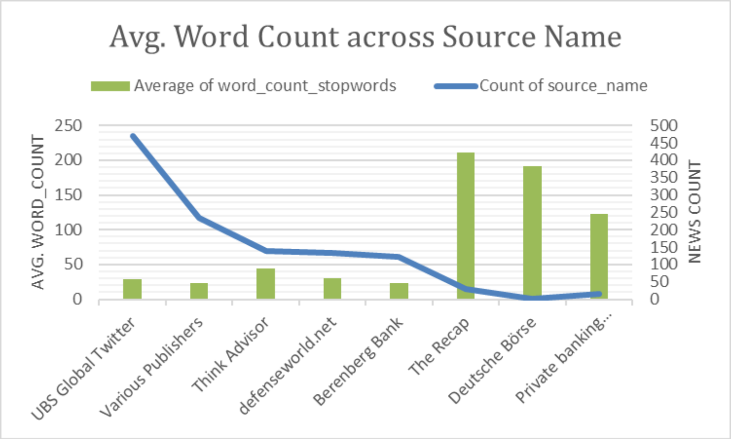
2. Dataset

Exploratory data analysis (3/3)



News Count across News Category

- Capital markets & stock analysis dominates with 47% of the dataset, recommending focused data extraction for a more balanced and comprehensive analysis across other news categories.



Avg. Word Count across Source Name

- UBS and ThinkAdvisor lead in news count, yet exhibit lower average word count per news article. In contrast, The Recap and Deutsche Börse show fewer news counts but deliver more extensive content.

Method

3. Method

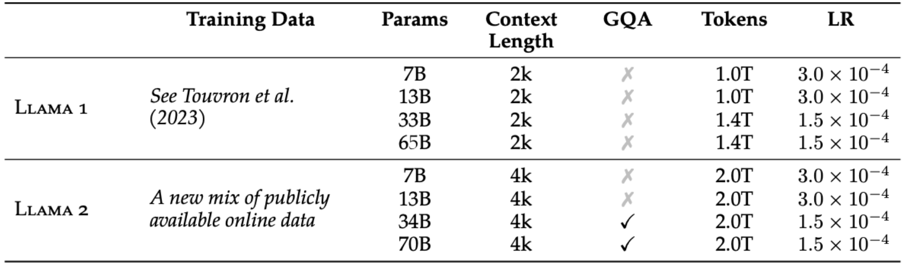
Model selection

Llama2 family of LLMs

- State of the art: large language models introduced by Meta
- Available in parameter counts ranging from 7 billion to 70 billion
- Requires GPUs for inferencing

We infer on Llama2-7b chat version; advantages:

- Open source – good community support
- Fine tuning is not computationally expensive
- Context window = 4096 tokens (i.e. ~ 3000 words)



	Training Data	Params	Context Length	GQA	Tokens	LR
Llama 1	*See Touvron et al. (2023)*	7B	2k	✗	1.0T	3.0×10^{-4}
		13B	2k	✗	1.0T	3.0×10^{-4}
		33B	2k	✗	1.4T	1.5×10^{-4}
		65B	2k	✗	1.4T	1.5×10^{-4}
Llama 2	*A new mix of publicly available online data*	7B	4k	✗	2.0T	3.0×10^{-4}
		13B	4k	✗	2.0T	3.0×10^{-4}
		34B	4k	✓	2.0T	1.5×10^{-4}
		70B	4k	✓	2.0T	1.5×10^{-4}

Table 1: Llama 2 family of models. Token counts refer to pretraining data only. All models are trained with a global batch-size of 4M tokens. Bigger models — 34B and 70B — use Grouped-Query Attention (GQA) for improved inference scalability.

https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/

3. Method

Extracting company names

LLMs recognize named entities

- LLMs have inherent capabilities to recognize named entities.*

- We leverage the power of *Llama-2-7B-Chat model* to enhance accuracy in discerning and isolating company names within the vast expanse of news content.

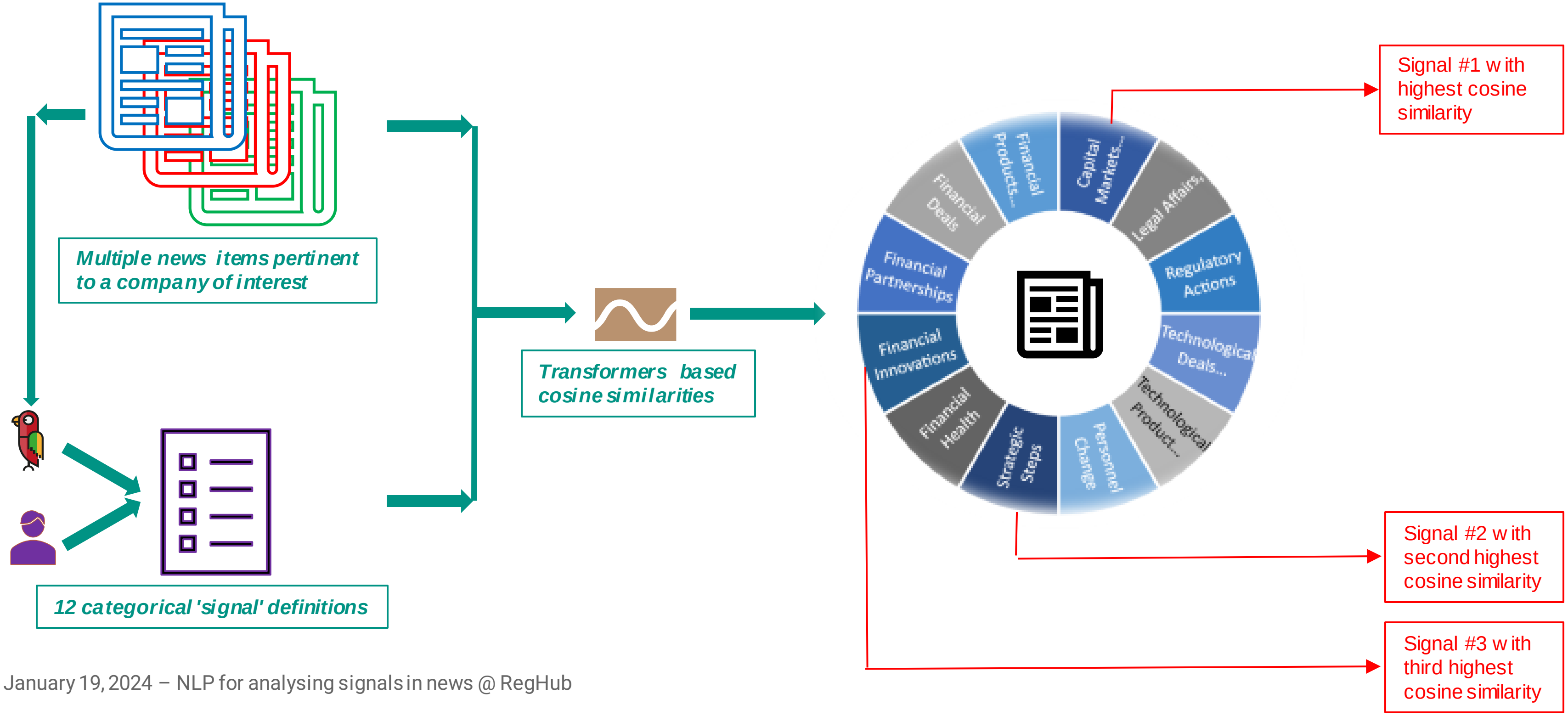
- Companies appear as ***subject*** and ***object*** within news items. LLMs have capabilities to identify the occurrence of these *noun* forms separately. This is essential for building better ***coherent contexts*** later in the project.

* https://www.clarifai.com/blog/do-llms-reign-supreme-in-few-shot-ner

BARCLAYS UBS BERENBERG PRIVATBANKIERS SEIT 1590 Deutsche Bank COMMERZBANK

3. Method

Spotting signals within news (1/2)



Multiple news items pertinent to a company of interest

12 categorical 'signal' definitions

Transformers based cosine similarities

Signal #1 with highest cosine similarity

Signal #2 with second highest cosine similarity

Signal #3 with third highest cosine similarity

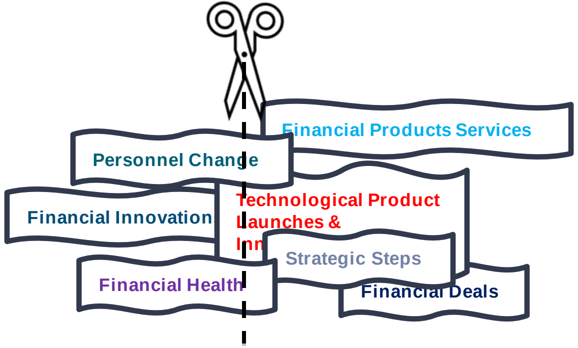
3. Method

Spotting signals within news (2/2)

Comparing news similarities with categorical signals

Idea: *Sentence similarities can classify news*.

○ Llama2 news summaries along with Reghub management's human expertise equipped us with 12 dimensions to categorize news content, encompassing aspects from "capital markets" and "financial health" to "technological innovations". Our task involved assigning relevant dimensions to each news item for comprehensive analysis.



Sentence embeddings in action

Implementation:

□ Leveraging the diversity in news dataset with the help of llama2, we came up with 12 definitions of categorical signals in agreement with RegHub. These signals cover a universe of different business-related news items.

□ Then we transformed (a) news content and (b) signal descriptions into embeddings using the Sentence-Transformers model (all-MiniLM-L6-v2 variant)

□ Using cosine similarity, we categorized news articles into four distinct signals based on their relevance to the embedded dimension descriptions.

□ The resulting signals range from Signal 1, representing the most relevant articles, to Signal 4, indicating articles with comparatively lower relevance, offering a nuanced and efficient classification system.

3. Method

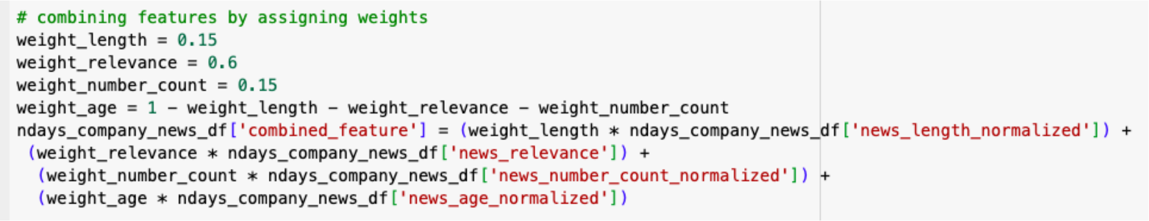
Ranking news items

Quantifying priorities of news items

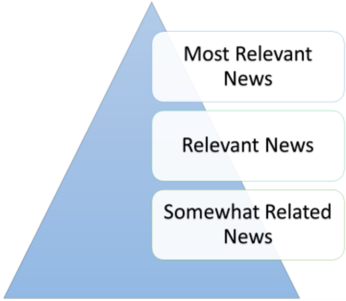
- **<u>Idea</u>**: *Numerous factors can help quantify the priorities of news items*.

- These may include, amongst others, the following:-
  - ***length of the news item*****:** higher length = better rank
  - ***news relevance*****:** more similarity to a signal definition = better rank
  - ***age of the news item*****:** younger, newer news = better rank
  - ***'number count' in the news item*****:** more numbers discussed in news = better rank

- ***<u>Implementation</u>***:
  - Filter news items on **desired company name** and **desired signal of relevance**.
  - Calculate **length, cosine similarity, age and number count** of each news items. Add more features/factors if necessary.
  - **Normalize** the calculated values for the filtered dataset.
  - Assign weights and calculate a proxy for a **combined-features score.**
  - Rank news items based on the 'combined-features score' evaluated above.



```
# combining features by assigning weights
weight_length = 0.15
weight_relevance = 0.6
weight_number_count = 0.15
weight_age = 1 - weight_length - weight_relevance - weight_number_count
ndays_company_news_df['combined_feature'] = (weight_length * ndays_company_news_df['news_length_normalized']) +
 (weight_relevance * ndays_company_news_df['news_relevance']) +
  (weight_number_count * ndays_company_news_df['news_number_count_normalized']) +
  (weight_age * ndays_company_news_df['news_age_normalized'])
```



3. Method

Unsupervised topic modelling

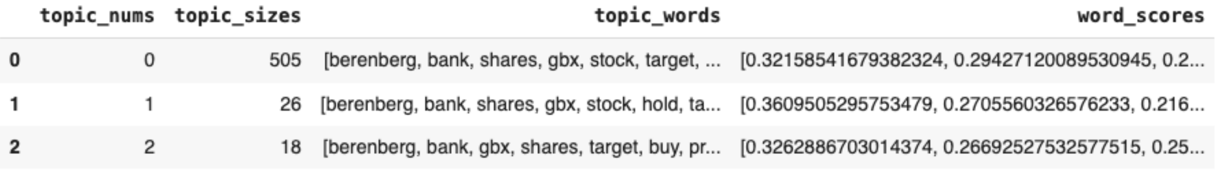
Unsupervised topic modelling

- **<u>Idea</u>**: *Use 'topic modeling' to discover latent semantic structure, i.e. topics, in a large collection of news documents*.

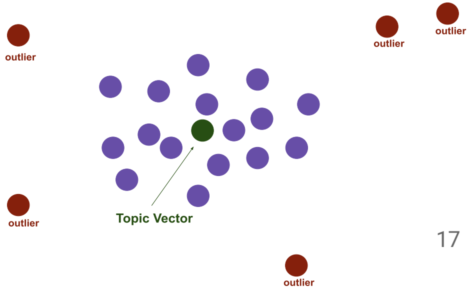
- Drawbacks of classical methods like **Latent Dirichlet Allocation**:
  - These often require the number of topics to be known, custom stop-word lists, stemming, and lemmatization.
  - Additionally, these methods rely on bag-of-words representation of documents which ignore the ordering and semantics of words.

- Top2Vec: We use instead distributed representations of documents and words due to their ability to capture semantics of words and documents.
  - It leverages joint document and word semantic embedding to find topic vectors.
  - The resulting topic vectors are jointly embedded with the document and word vectors with the distance between them representing semantic similarity.

- ***Implementation:***
  - News items were filtered on **desired company name.**
  - Unsupervised topic modelling identified **3 topics** within filtered news items.
  - Top2Vec performs better with larger text corpuses. Topic overlaps observed in our implementation.

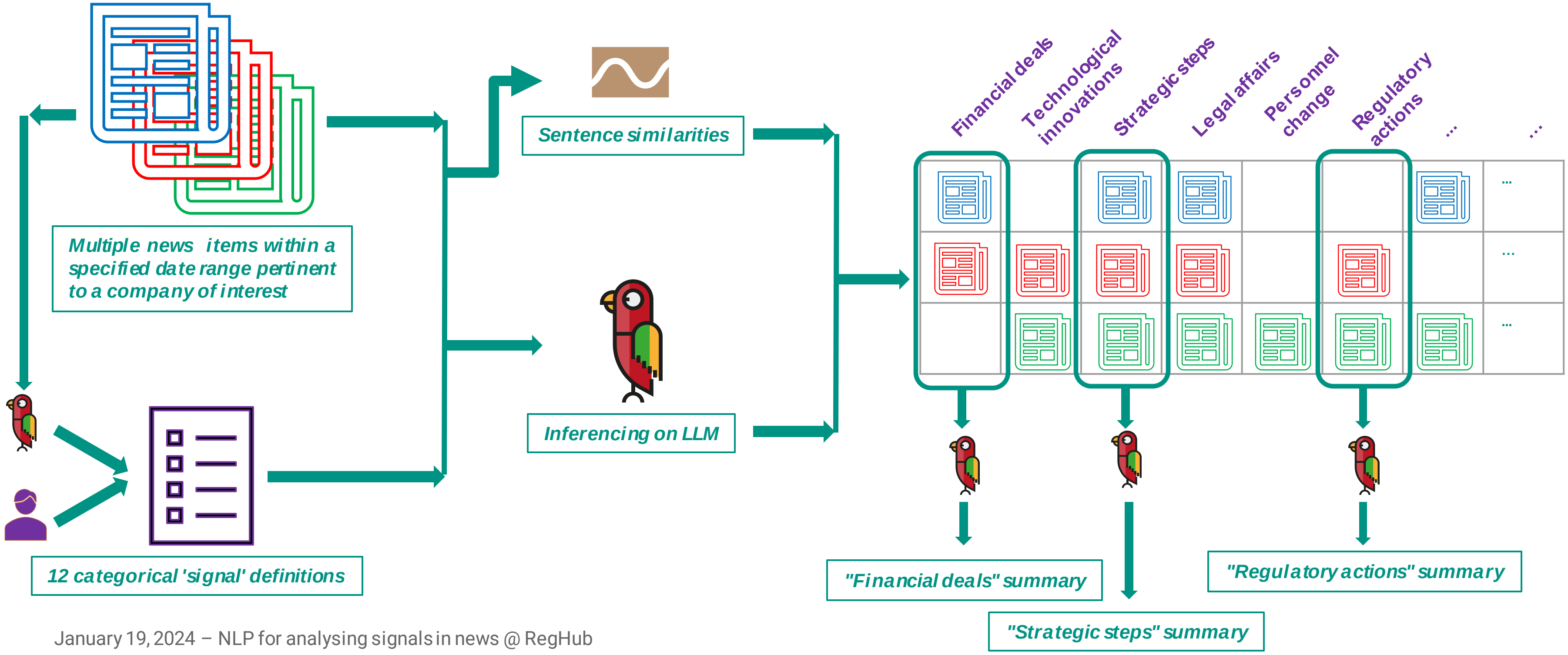


	topic_nums	topic_sizes	topic_words	word_scores
0	0	505	[berenberg, bank, shares, gbx, stock, target, ...	[0.32158541679382324, 0.29427120089530945, 0.2...
1	1	26	[berenberg, bank, shares, gbx, stock, hold, ta...	[0.3609505295753479, 0.2705560326576233, 0.216...
2	2	18	[berenberg, bank, gbx, shares, target, buy, pr...	[0.3262886703014374, 0.26692527532577515, 0.25...



3. Method

Capturing 'temporal' progress in news signals (1/3)



Multiple news items within a specified date range pertinent to a company of interest

12 categorical 'signal' definitions

Sentence similarities

Inferencing on LLM

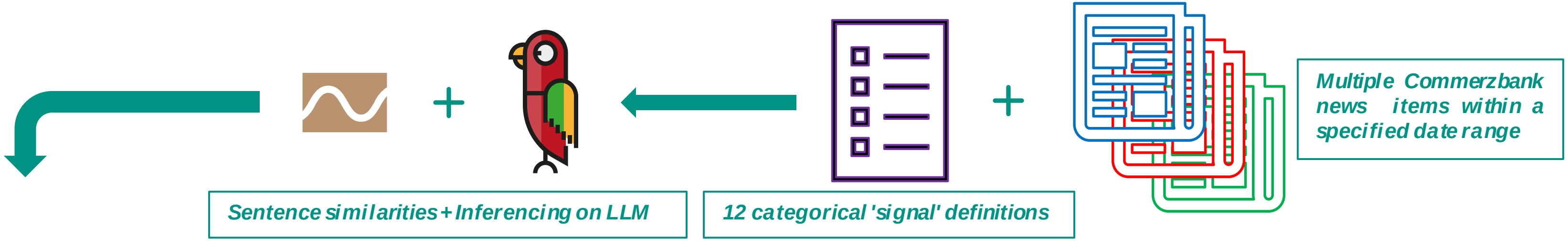
"Financial deals" summary

"Strategic steps" summary

"Regulatory actions" summary

3. Method

Capturing 'temporal' progress in news signals (2/3)



Multiple Commerzbank news items within a specified date range

12 categorical 'signal' definitions

Sentence similarities + Inferencing on LLM

Personnel change

- *Commerzbank is undergoing significant changes in its personnel structure. The appointment of Bernd Spalt as the new Chief Risk Officer is a notable change, as he brings experience from his previous role as CEO of Erste Group Bank. This suggests that the company is prioritizing risk management and financial stability.*

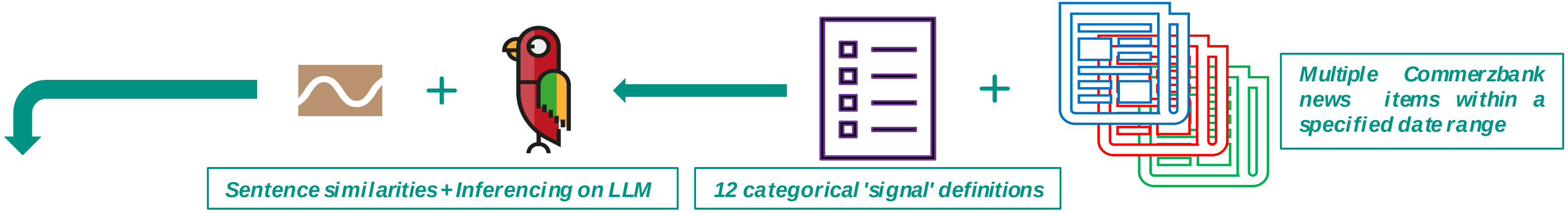
Legal affairs

- *Lawsuit against six major German banks for their alleged role in facilitating money laundering, seeking to recover funds lost and impose punitive damages.*
- *Commerzbank is suing accounting firm EY over the 200 million euros in losses incurred in the collapse of Wirecard.*

Regulatory actions

- The company has appointed a new Chief Risk Officer, Bernd Spalt, to enhance risk management operations and improve financial stability.
- Commerzbank is taking legal action against EY over the 200 million euros in losses incurred in the collapse of Wirecard, which was involved in regulatory issues and faced consequences from regulatory bodies.
- The Royal Court of Appeal (RCA) has filed a lawsuit against six major German banks, including Commerzbank, for their alleged role in facilitating money laundering.

Capturing 'temporal' progress in news signals (3/3)



Multiple Commerzbank news items within a specified date range

12 categorical 'signal' definitions

Sentence similarities + Inferencing on LLM

Strategic steps

1. *Commerzbank CEO Manfred Knof has warned of the need for structural changes in Germany to address the growing appeal of right-wing political groups.*
2. *The company is planning to invest in its workforce to meet the demands of digitalization and innovation.*
3. *Commerzbank is planning to cut jobs in its wealth management division in Asia, reflecting challenges faced by global banks in the region.*
4. *The European Central Bank (ECB) has put a hold on its plan to increase interest rates due to economic challenges and uncertainties faced by the Eurozone.*
5. *Commerzbank played a role in the issuance of a bond by Sixt SE, which is a significant corporate move.*
6. *Several equities research analysts have recently issued reports on the company, including a report by Bloomberg Intelligence that suggests a merger could provide significant cost savings.*
7. *UBS and Credit Suisse are potential merger partners, with a merger potentially positioning Iqbal Khan as the eventual successor to CEO Ralph Hamers.*
8. *The sale of a stake of over five percent in Commerzbank and Deutsche Bank, totaling about 1.75 billion euros, resulted in a sharp fall in the shares of the top two German lenders.*

Technological product innovation

- Commerzbank has launched a new digital platform for its corporate clients, which offers a range of services including account management, payment transactions, and cash management.
  The platform, called "Commerzbank One," aims to provide a more streamlined and efficient experience for clients, and has already seen significant adoption since its launch.
  Additionally, the company has announced plans to invest €1 billion in its digital transformation over the next three years, with a focus on developing new digital products and services.

Results

4. Results

Cost-compute-latency estimate analysis

activity*	latency* (ms)	compute** (TFLOPS)	cost** ($)	remarks
Extracting company names	100	0.02	0.005	Utilizing llama2-7b for entity extraction. Latency and compute are relatively low due to the model's efficiency in recognizing entities. Cost based on minimal resources.
Spotting top 4 signals within news	400	0.1	0.02	Calculating cosine similarity with signal definitions. Moderate latency and compute due to similarity calculations. Cost slightly higher due to increased computation.
Ranking news items	50	0.005	-	Straightforward arithmetic operations for ranking. Very low latency, compute, and cost.
Unsupervised topic modelling	200	0.5	0.01	Using Top2Vec for unsupervised topic modelling. Moderate latency and compute. Cost includes resources for topic modelling.
Capturing temporal progress in news signals	200	0.2	0.005	Sum of spotting signals and LLM inferencing for temporal decomposition. Moderate latency and compute. Higher cost due to increased complexity.
Overall (per news item basis)	**950**	**0.825**	**0.04**	**Aggregated sum of all activities. Total latency, compute, and cost.**

* inferencing performed on standard T4 GPUs available on Google colab platform

** estimated in proportion of average tokens; referring to the Dell Technologies whitepaper (mentioned under key references)

Challenges and learnings

Challenges

- **Unstructured nature of textual dataset:**
  Language datasets are let alone entropic, and for our case, the dataset comprised of news scraped from multiple web sources.

- **Computational constraints:**
  Limited computational power constrained us to employ lower models, restricted to 4-bit training.
  Temporal information capture was focused on a single company at a time due to computational challenges.
  The project was conducted on Google Colab, utilizing a T4 GPU with a daily limit of 15GB.

- **Small dataset limitations**:
  The project encountered challenges stemming from a relatively small dataset, impacting the depth and breadth of analysis. Visibility of raw data and initial processing methodology is limited. Many a times, news content is present as a link instead of news text.

- **Manpower limitations for reinforcement learning**:
  The unavailability of ample human feedback data hindered the implementation of reinforcement learning on our transformers. Additional manpower could have facilitated the creation of a more robust human feedback dataset for refining model performance.

Learnings

- **Real-world problem solving by amalgamating classic NLP with the modern LLM realm**:
  Implementation of classic NLP methods along with modern LLM methods provided valuable insights into addressing real-world challenges. Amalgamating these two, we successfully captured temporal signals within textual news.

- **Cost-quality tradeoff**:
  Recognizing the pivotal role of cost in model enhancement, we learned that increasing accuracy often comes with additional expenses – either in the form of time or money. The project underscored the delicate balance between investing in model improvements and managing the overall time, emphasizing the importance of a strategic tradeoff.

- **Exposure to present day computational resources and complexities:**
  Working on this project exposed us to modern day computational optimization practices as well as the challenges that follow. *Quantization methods* and *cloud based computing* were a part of this learning.

Summary and future outlook

Summary

- Categorical signals can be extracted from temporal unstructured textual data by strategically employing classic NLP and modern LLM methods.
- Manual annotation is tedious and can be replaced by AI methods at lower time-cost and comparable accuracies.
- Topic modelling works better with even larger datasets.
- News classification can be achieved by comparing distances (cosine sim.) between embeddings of news items with categorical signal definitions.
- Categorization accuracy is hard to measure, as several categories are often applicable to one News (at least for the Reghub news dataset example).
- A lot of data cleaning can be required if the structure of News dataset changes in any way.

Next steps

- Incoming news should be stored in a consistent format, to avoid high costs for data cleaning.
- More categories should be tested and included than in this pilot.
- Bigger models (>7b parameters) will outperform the results achieved by us.
- Larger training datasets for fine-tuning, ample human feedback for RLHF and advance computational resources will contribute towards higher accuracies.

Key references and Github code repository

Key references

- Angelov, Dimo. "Top2vec: Distributed representations of topics." arXiv preprint arXiv:2008.09470 (2020).
- Dell Technologies. "Llama 2: Inferencing on a Single GPU." Whitepaper, September 2023.
- Liu, Huicheng. "Leveraging financial news for stock trend prediction with attention-based recurrent neural network." arXiv preprint arXiv:1811.06173 (2018).
- Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
- Wang, Wenhui, et al. "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers." Advances in Neural Information Processing Systems 33 (2020): 5776-5788.
- Zhang, Yunyi, et al. "Unsupervised Key Event Detection from Massive Text Corpus." (2022).

https://github.com/neelblabla/nlp_for_analysing_news_signals