ELECTRONIC ASSIGNMENT COVERSHEET

Student Number:	34445873
Student Name:	Chew Jian Yue

Unit Code:	ICT606
Unit name:	Machine Learning
Enrolment mode:	External
Date:	22 July 2023
Assignment number:	2
Assignment name:	Project
Tutor:	Mr. Johnson Ang

Student's Declaration:

1. Except where indicated, the work I am submitting in this assignment is my own work and has not been submitted for assessment in another unit.
2. This submission complies with Murdoch University's academic integrity commitments. I am aware that information about plagiarism and associated penalties can be found at http://our.murdoch.edu.au/Educational-technologies/Academic-integrity/. If I have any doubts or queries about this, I am further aware that I can contact my Unit Coordinator prior to submitting the assignment.
3. I acknowledge that the assessor of this assignment may, for the purpose of assessing this assignment:
   a. reproduce this assignment and provide a copy to another academic staff member; and/or
   b. submit a copy of this assignment to a plagiarism-checking service. This web-based service may retain a copy of this work for the sole purpose of subsequent plagiarism checking, but has a legal agreement with the University that it will not share or reproduce it in any form.
4. I have retained a copy of this assignment.
5. I will retain a copy of the notification of receipt of this assignment. If you have not received a receipt within three days, please check with your Unit Coordinator.

I am aware that I am making this declaration by submitting this document electronically and by using my Murdoch ID and password it is deemed equivalent to executing this declaration with my written signature.
Optional Comments to Tutor: *e.g. If this is a group assignment, list group members here* Word count is within 1100 + 10%

*If you can, please insert this completed form into the body of **each** assignment you submit. Follow the instructions in the Unit Information and Learning Guide about how to submit your file(s) and how to name them, so the Unit Coordinator knows whose work it is.*

Start your assignment on the next page.

Assignment 2

Topic modelling on financial news headlines with effect on stock prices

Chew Jian Yue (34445873)

Table of Contents

Abstract

Topic modelling is helpful in identifying latent subjects in corpus documents. In particular, short text clustering, prominent in social networking and news headlines, is characterised by high sparsity, dimensionality and volume [1].

Introduction

The purposes of this paper are:

- Identifying topics within short financial news headlines relating to selected stocks
    - Kroger stock: A publicly listed American retail company that operates supermarkets and multi-department stores[2].
- News headlines influencing 1-day stock price movements as derivation to market sentiment.
- Kroger management could identify media publicity topics; competitor analysis.
- How topic models changed over time.

According to efficient-market hypothesis theory, semi-strong efficient markets react to public information instantaneously, reflecting the information in market prices. Hence, positive news should reflect an increase in share price, vice-versa.

Data collection and EDA

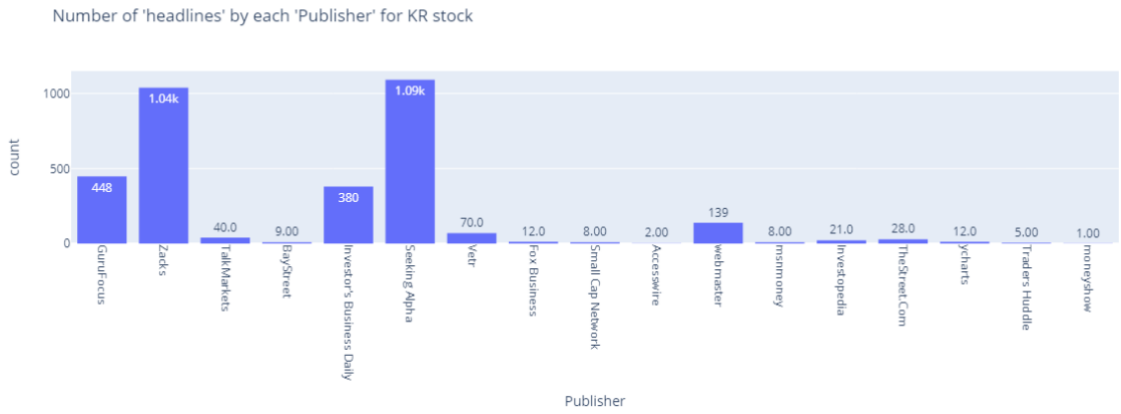


Figure 1

Data sources used are static Kaggle news headline dataset.

```
df_headlines = apply_dataset_df(df_headlines)
# df_headlines["Date"] = pd.to_datetime(df_headlines["Date"], utc = True).dt.date

df_headlines = df_headlines.query("Stock == 'KR'") # keep only headlines of KR stock

df_headlines.head()
```

✓ 1.7s

	Unnamed: 0.1	Unnamed: 0	Headline	url	publisher	Date	Stock
956318	956318	958554	Kroger First Quarter Conference Call with Inve...	http://www.gurufocus.com/news/1155641/kroger-f...	GuruFocus	2020-06-04	KR
956319	956319	958555	Nordstrom (JWN) Q1 Loss Wider Than Expected, S...	http://www.zacks.com/stock/news/951775/nordstr...	Zacks	2020-05-29	KR
956320	956320	958556	Looking for Dividend Growth? Here Are 5 Solid ...	http://www.zacks.com/stock/news/951451/looking...	Zacks	2020-05-29	KR
956321	956321	958557	Kroger: Capital Preservation and More at a Goo...	http://www.gurufocus.com/news/1149842/kroger-c...	GuruFocus	2020-05-28	KR
956322	956322	958558	SpartanNash (SPTN) Q1 Earnings Surpass Estimat...	http://www.zacks.com/stock/news/950122/spartan...	Zacks	2020-05-28	KR

Data sources	Data description	Reasons for selection

Kaggle Daily Financial News for 6000+ Stocks from 2009-2020
 	
(*raw_partner_headlines.csv*)
	Static news headlines collected between 2009 to 2020 consisting of around 4 million articles across about 6000 stocks.
	
	Data engineer/author collated headlines from Benzinga.com through bot web scraping [3].
		Dataset has:
		
		• Usability score of 10 on Kaggle, which assesses on:
		  ○ Completeness: clear, detailed documentation on data attributes
		  ○ Compatibility: available in plaintext comma separated values (CSV) format, readable by Python libraries

	Includes the following attributes/features: 1. Article headlines 2. URL to article 3. Publisher 4. Date and timestamp 5. Stock ticker	○ Credibility: indication of original news sources including Seeking Alpha and GuruFocus • Belongs to public domain, consists of news headlines rather recently (3 years ago) and allows for longitudinal time analysis. News data available almost every trading day between the indicated period • Large enough dataset (~400 megabytes) for flexibility in reliable analysis
Yahoo! Finance Stock Market Data (Kruger share price)	Provides rich consolidated resource of financial market data and tools to assess investments [4]	• Used as a self-supervised indicator for "positive", "neutral" and "negative", correlating impact of news sentiment on stock prices • Many attributes can be extracted from easy-to-use Python library yfinance that interfaces with public Yahoo! Finance API used in their website • Provides extensive historical closing stock prices of my choice in pandas DataFrame datatype, allowing for easy integration by joining with other data sources

```
Value counts of stocks, sorted by largest number of 'headlines':
KR       3314
GXC      3238
PGJ      3082
YINN     3027
JPM      2873
         ...
SLMAP       1
CADT        1
CVB         1
WMT         1
JBK         1
Name: stock, Length: 6552, dtype: int64
Stock symbol with largest number of 'headlines': KR
```

Headlines related to Kroger are specifically used due to its frequency of headlines.

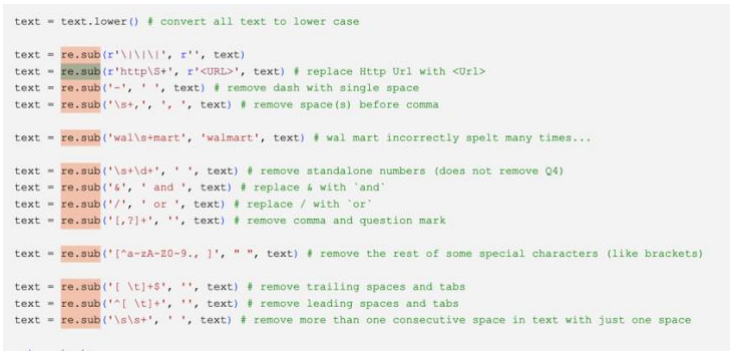
These short and concise headlines are inexpensive to collect in large volumes through RSS feeds, web scraping, APIs and mimics character-limited tweets during deployment and production. For some tasks, including tokenizer models, token max length is critical (BERT: 512).

Data pre-processing

The first run of topic modelling identified issues when evaluating generated results when eye-balling. For example, additional stopwords are identified beyond those listed in NLTK library.

Problems	**Considerations with applied solutions**
Many frequent and common words	Extended with custom stopwords list Use of c-TD-IDF during modelling

Lack of diverse topics	`representation_model = MaximalMarginalRelevance(diversity=0.8) # towards 1 = more diverse`
	`representation_models.append(representation_model)`
	Use MMR[i] representation and changing hyper-parameter "diversity" during modelling

Excessive topics identified	Studying the visualisations and defining hyper-parameter "nr_topics" with "reduce_topics" method to reduce total number of topics with c-TD-IDF calculation
Irregular punctuations, Special characters	 Applied Regex substitutions and apply across headlines series Special characters when removed may results in semantic meaning loss (e.g., full stop represents end of sentence), while others do not carry meaning and only results in worse model performance "walmart) " is similar to "walmart".

Misspelt common words	e.g. Replacing “wal mart” with “walmart” to improve model performance through accurately calculating word frequencies
Letters with numbers	“q4” refers to quarter 4 and may have some relationship to earnings. This is important word and should not be removed in this context.

```
CPU times: user 794 ms, sys: 402 µs, total: 794 ms
Wall time: 798 ms
```

	Unnamed: 0.1	Headline	publisher	Date	Stock	Close	Close_1d_percent_change	Close_1d_diff	Change from previous 1d close
0	956318	first quarter conference call investors	GuruFocus	2020-06-04	KR	31.131811	0.013158	0.404310	Up
1	956319	nordstrom jwn q1 loss wider expected sales	Zacks	2020-05-29	KR	30.671085	0.017467	0.526539	Up
2	956320	looking dividend growth solid picks	Zacks	2020-05-29	KR	30.671085	0.017467	0.526539	Up
3	956321	capital preservation good price	GuruFocus	2020-05-28	KR	30.144547	-0.036659	-1.147108	Down
4	956322	spartannash sptn q1 earnings surpass estimates...	Zacks	2020-05-28	KR	30.144547	-0.036659	-1.147108	Down
...	...	...	...	...	...	...	...	...	...
3309	959627	forecast could fall short	webmaster	2010-03-09	KR	8.776717	-0.024017	-0.215979	Down
3310	959628	cisco systems tuesday headlines	webmaster	2010-03-09	KR	8.776717	-0.024017	-0.215979	Down
3311	959629	rocket stocks week	webmaster	2010-03-08	KR	8.992696	0.007036	0.062830	Up
3312	959630	cramer mad money recap next week game plan update	webmaster	2010-03-05	KR	8.929866	0.002203	0.019633	Up
3313	959631	alternative investing new normal	webmaster	2010-02-04	KR	8.272469	-0.016729	-0.140745	Down

3048 rows × 9 columns

Figure 2-Processed DataFrame

Figure 2 shows final transformed DataFrame by joining the stock with headlines.

Topic modelling building and explanation

	Advantages	Disadvantages
LDA	• Prior domain knowledge is not necessarily required • Finds coherent topics when correct hyperparameter tuning is applied • Can deal with sparse input • The number of topics is generally smaller than word-embedding based approaches; thus, it is easier to be interpreted • One document can contain several different topics (Mixed membership extraction) • Full generative models with multinominal distribution over topics are generated • Shows both adjectives and nouns within topics	• Detailed assumptions are required • Hyperparameters need to be tuned carefully • Results can easily produce overlapping topics as topics are soft clusters • Objective evaluation metrics are widely missing • The number of topics needs to be defined by the user(s) • Since the results are not deterministic, reliability and validity are not automatically ensured • Assumes that the topics are independent of each other; hence, only the frequency of the common occurrence of words is used • Word correlations are ignored, so no relationships between topics can be modeled
NMF	• Prior domain knowledge is not required • Supports mixed membership models; thus, one document can contain several topics • In contrast to LDA, which uses raw word frequencies, the term-document matrix can be weighted with TF-IDF • It proves to be computationally efficient and very scalable • Easy to implement	• Frequently delivers incoherent topics • The number of topics to be extracted must be defined by the user in advance • Implicit specification of probabilistic generative models
Top2Vec	• Supports hierarchical topic reduction • Allows for multilingual analysis • Automatically finds the number of topics • Creates jointly embedded word, document, and topic vectors • Contains built-in search functions (easy to go from topic to documents, search topics, etc.) • Can work on very large dataset sizes • It uses embeddings, so no preprocessing of the original data is needed	• The embedding approach might result in too many topics, requiring labor-intensive inspection of each topic • Generates many outliers • Not very suitable for small datasets (<1,000) • Each document is assigned to one topic • Objective evaluation metrics are missing
BERTopic	• High versatility and stability across domains • Allows for multilingual analysis • Supports topic modeling variations (guided topic modeling, dynamic topic modeling, or class-based topic modeling) • It uses embeddings, so no preprocessing of the original data is needed • Automatically finds the number of topics • Supports hierarchical topic reduction • Contains built-in search functions (easy to go from topic to documents, search topics, etc.) • Broader support of embedding models than Top2Vec	• The embedding approach might result in too many topics, requiring labor-intensive inspection of each topic • Generates many outliers • No topic distributions are generated within a single document; rather, each document is assigned to a single topic • Objective evaluation metrics are missing

Table 1–Comparison between modelling methods from[5]

BERTopic provides workflow pipeline's modularity leveraging on transformers and c-TF-IDF to create clusters for interpretable topics while keeping important words[6] with varied model representations and variations[7]. Unlike LDA, BERTopic (like Top2Vec) offers continuous instead of discrete modelling. Studies suggest BERTopic embedding models generate novel insights with topic reduction surpassing performance of Top2Vec using pretrained embeddings cover more overlapping topics[5], [8]. Topics produced by LDA were unintriguing although running on CPU. Embeddings capture textual semantic meanings while NMF's low capability. Table 1 summarises the benefits outweighing costs of using BERTopic.

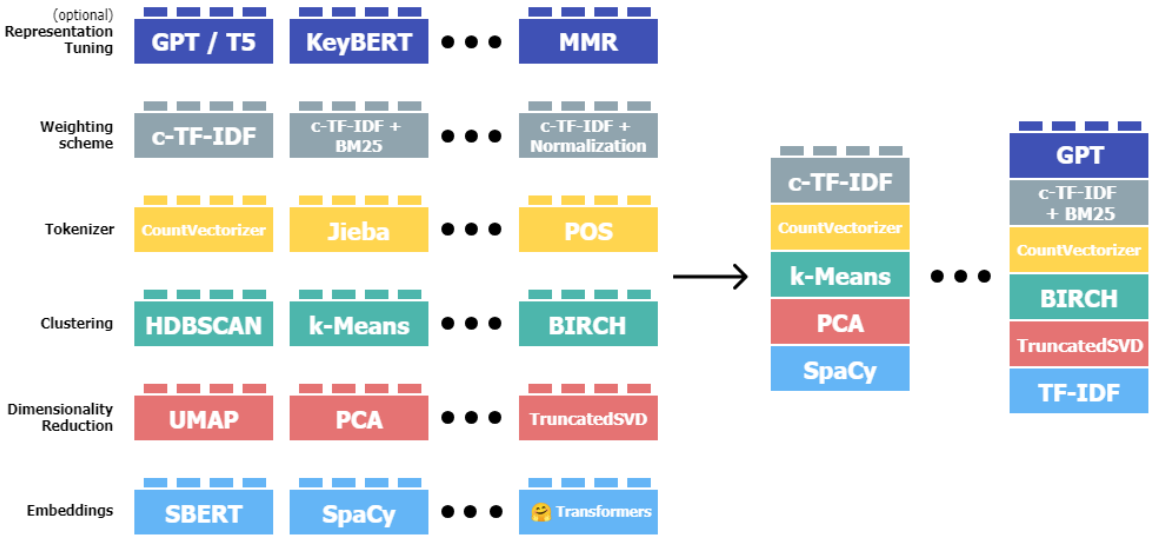


Figure 3-BERTopic pipeline[6]

Embeddings[5]

```
SentenceTransformer(
  (0): Transformer({'max_seq_length': 256, 'do_lower_case': False}) with Transformer model: BertModel
  (1): Pooling({'word_embedding_dimension': 384, 'pooling_mode_cls_token': False, 'pooling_mode_mean_tokens': True, 'pooling_mode_max_tokens': False, 'pooling_mode_mean_sqrt_len_tokens':
  (2): Normalize()
)
```

Headlines are in English, S-BERT "all-MiniLM-L6-v2" is used as good performing pre-trained sentence transformer model. *SentenceTransfomers* provides embeddings at sentence-level, providing semantic similarity are superior to BERT[9].

Dimensionality reduction

```
▼ UMAP
UMAP(metric='cosine', min_dist=0.0, n_components=5, random_state=0)
```

Since clustering algorithms are sensitive to "curse of dimensionality", making equidistant points difficult to cluster[9], dimensionality reduction is performed with UMAP[ii] using cosine similarity.

Clustering

```
▼ HDBSCAN
HDBSCAN(min_cluster_size=15, prediction_data=True)
```

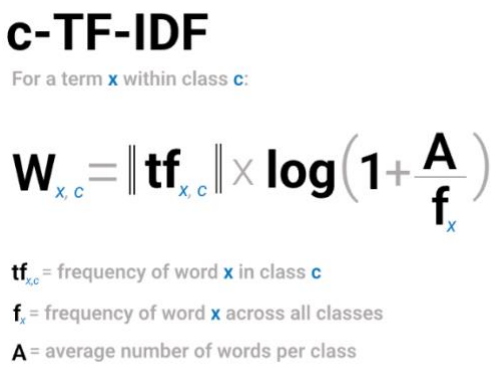
HDBSCAN, a hierarchical density-based clustering approach. Points furthest are marked as outliers and automatically reduce topics with cluster sizes at least 15.

Tokenizer

```
CountVectorizer
CountVectorizer(min_df=10, ngram_range=(1, 2), stop_words='english',
                tokenizer=<__main__.LemmaTokenizer object at 0x7f718e5d4970>)
```

Parameter	**Explanation**
tokenizer	Lemmatization callback function using NLTK.stem's WordNetLemmatizer reducing words to dictionary form with context consideration[10]
stop_words	Detect and filter stop words uncaught in pre-processing
ngram_range	Most words come as unigram and in pairs

c-TF-IDF




```
ClassTfidfTransformer
ClassTfidfTransformer(bm25_weighting=True, reduce_frequent_words=True)
```

Figure 4 - Class-based TF-IDF procedure

Used to extract topic representations, making one cluster differ from another, frequent words reduced automatically.

Evaluation

Intrinsic - Coherence

A group is coherent when words within a group corroborate with one another.

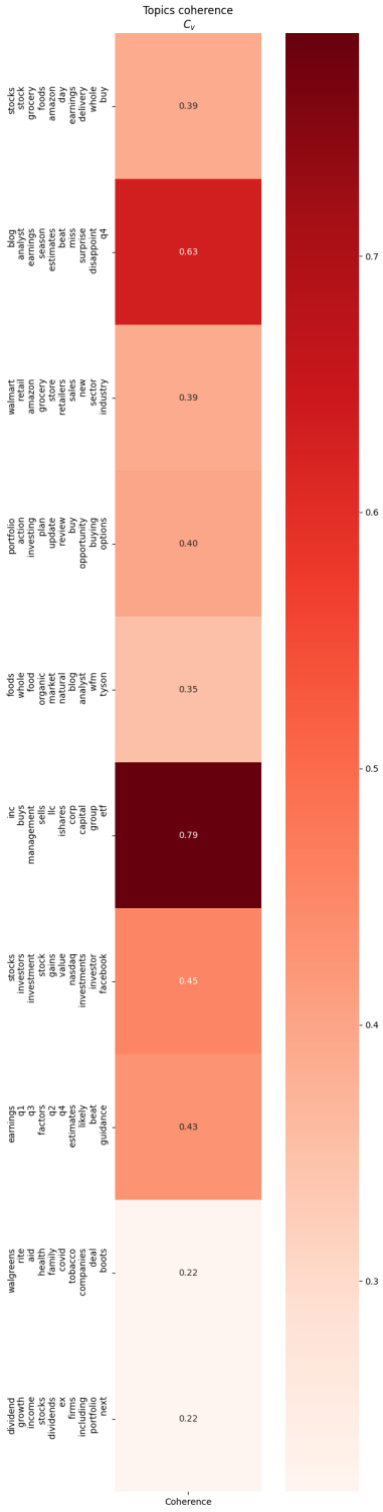
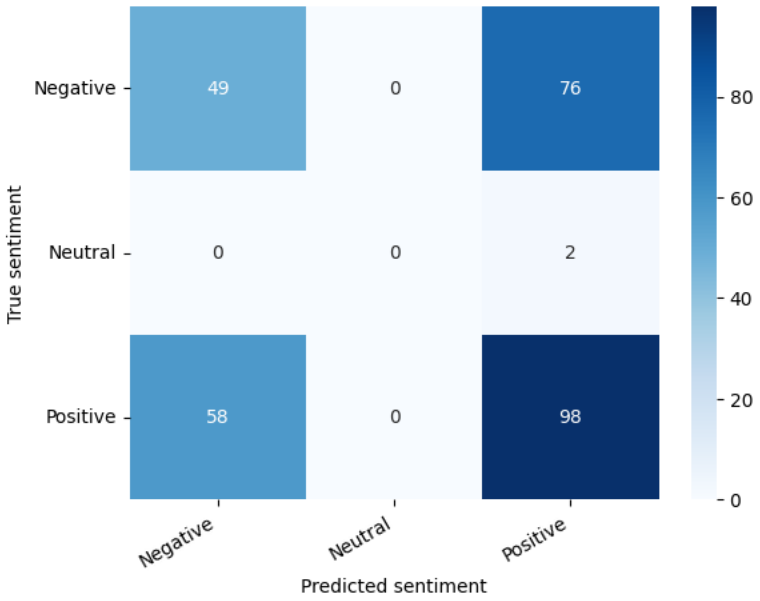
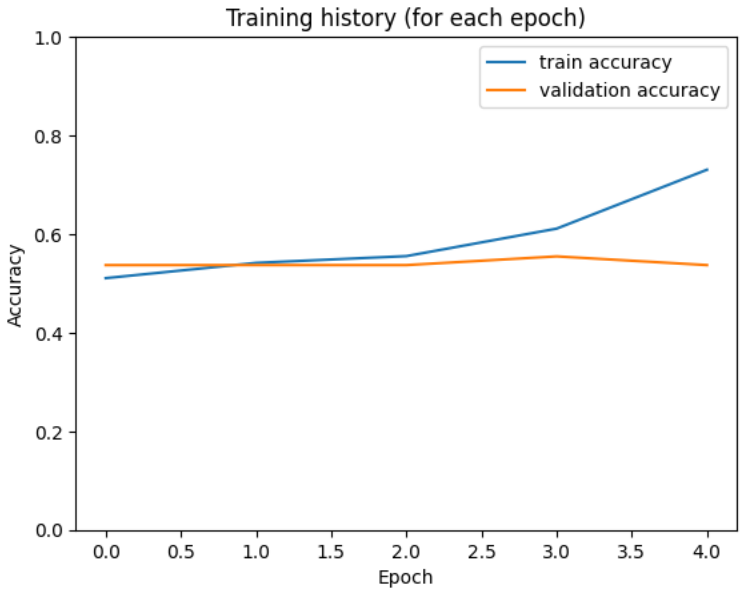


Figure 5-Coherence heatmap

The first ten topics are extracted and coherence scores are calculated. The more coherent topic words: [analyst,estimates,beat,disappoint,q4,…] with 0.63 and [buys,sells,etf,management,…] with 0.79 corresponds to human evaluator's.

Extrinsic-Sentiment performance





The sentiment model predicts “positive” usually, and could be over-fitting as training epochs increases.

Topic modelling results, visualization, and interpretation

Figure 6-Wordcloud

Most headlines have words “amazon”, a competitor, “store”, “grocery” which relates to Kroger’s business and “earnings”, “q3” which relates to announcements of earnings.

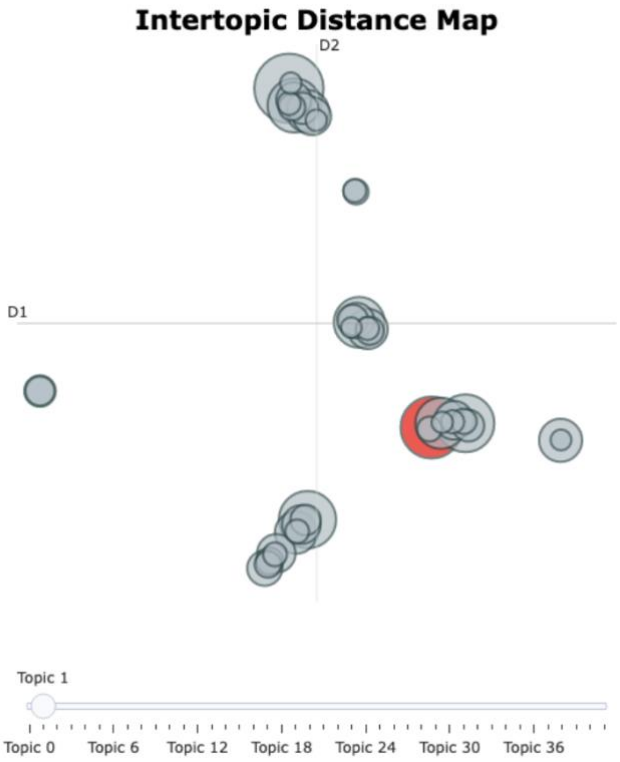


Figure 7-Intertopic Map

The 2-D representation shows topics clustering and extent of topic differences.

```
print(topics[104])     "topics" is not defined
print(df["Headline"].iloc[104])     "iloc": Unknown word.
topic_model.visualize_distribution(probs[104], min_probability=0.015)     "
```

✓ 0.0s

```
2
amazon whole foods arranges special hours senior citizens
```

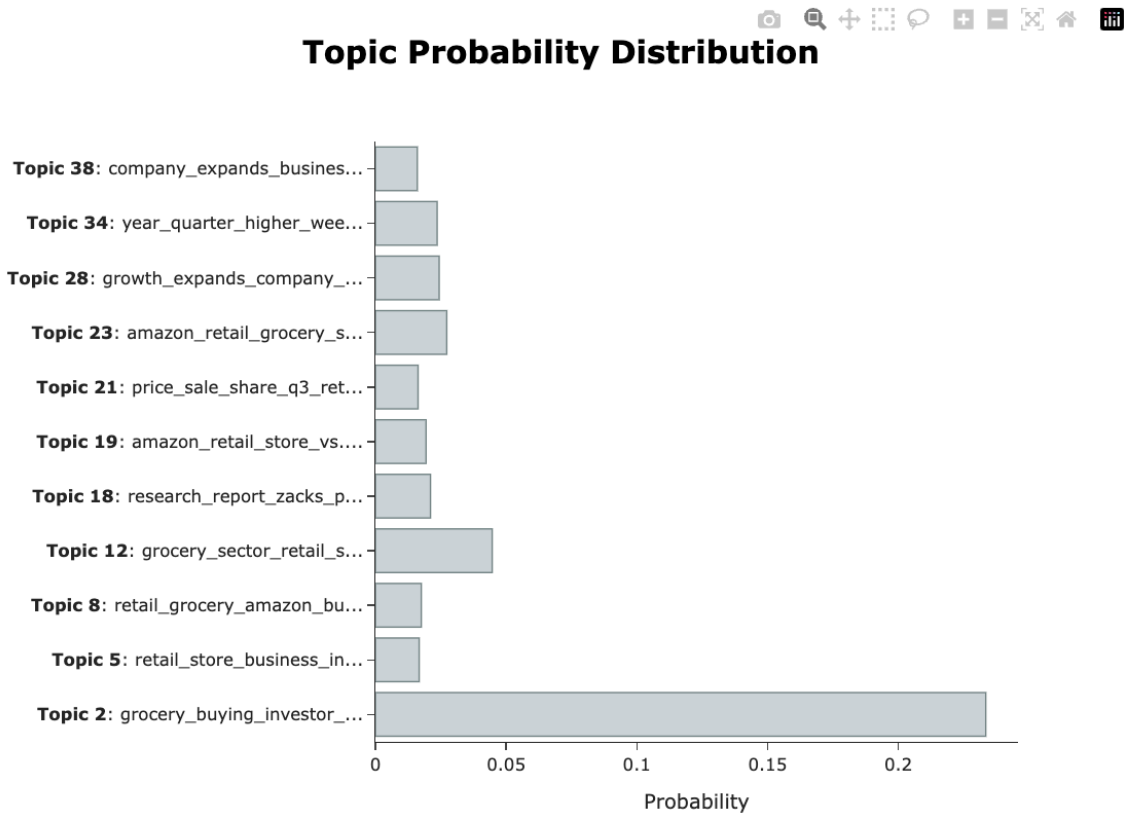


Figure 8-Example of headline topic

Headline #105 highly likely belongs to topic 2.The headline could mean competition is becoming steeper for Kroger as competitors are amping their CSR[iii], which is hardly captured in the topic model.

```
print(df["Headline"][104])
```

✓ 0.0s

```
combat coronavirus major retailers add seniors shopping hours
```

```
# To understand each token's contribution to a specific topic: calculate topic distributions on a token level and visualise

## Calculate the topic distributions on a token-level
topic_distr, topic_token_distr = topic_model.approximate_distribution(df["Headline"], calculate_tokens=True)    "topic_model"

## Visualize the token-level distributions
df_vad = topic_model.visualize_approximate_distribution(df["Headline"][104], topic_token_distr[104])    "topic_model" is not
df_vad
```

✓ 1.8s

```
100%|██████████| 4/4 [00:01<00:00,  2.17it/s]
```

	combat	coronavirus	major	retailer	add	senior	shopping	hour
2_grocery_buying_investor_company	0.238	0.492	0.745	0.745	0.507	0.253		
12_grocery_sector_retail_store	0.112	0.238	0.364	0.364	0.251	0.126		
18_research_report_zacks_post	0.164	0.348	0.532	0.532	0.368	0.184		
23_amazon_retail_grocery_store	0.126	0.126	0.126	0.126				
28_growth_expands_company_long term	0.122	0.248	0.374	0.374	0.252	0.126		

Figure 9-Transparency of words' topics

The token's contribution to topics shows "retailer" is an contributing factor to topic 2.

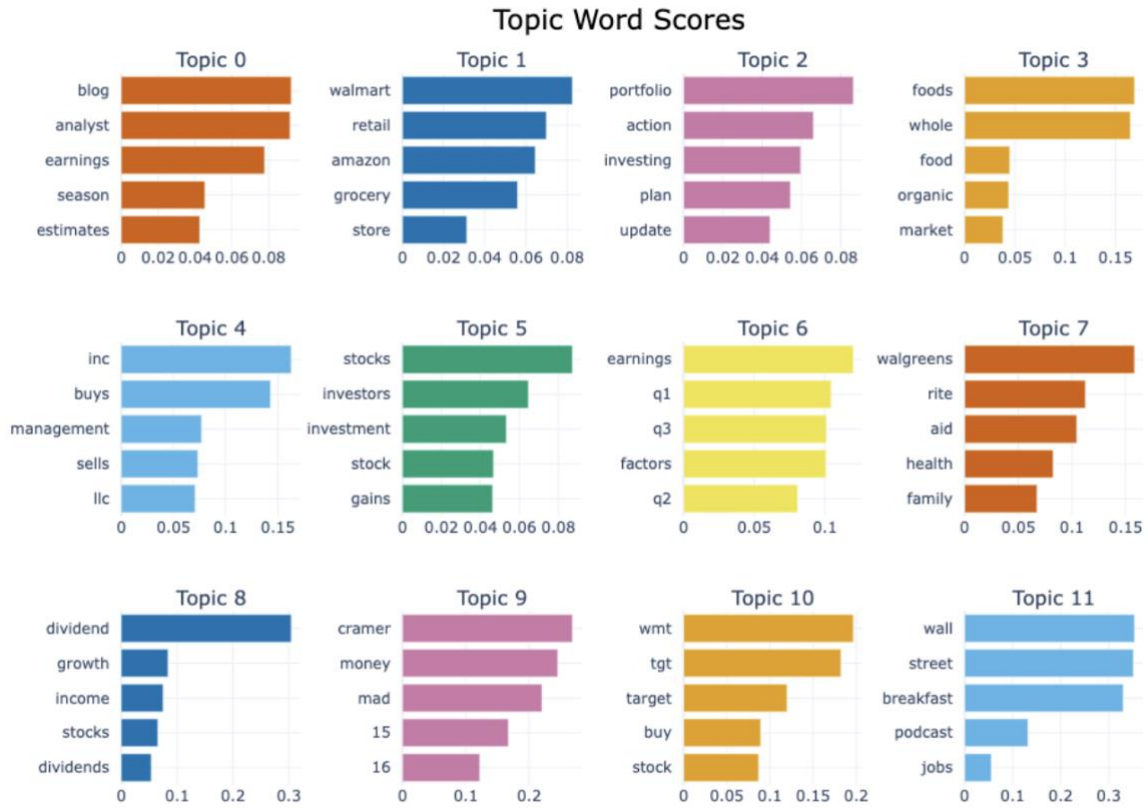


Figure 10-Topic word scores

Figure 10 shows probability of terms in top topics (from c-TF-IDF scores). For instance, topics 1,7 talks about Kroger competitors, 5,6,8 talks about financial performance/investment and 3 talks about its business.

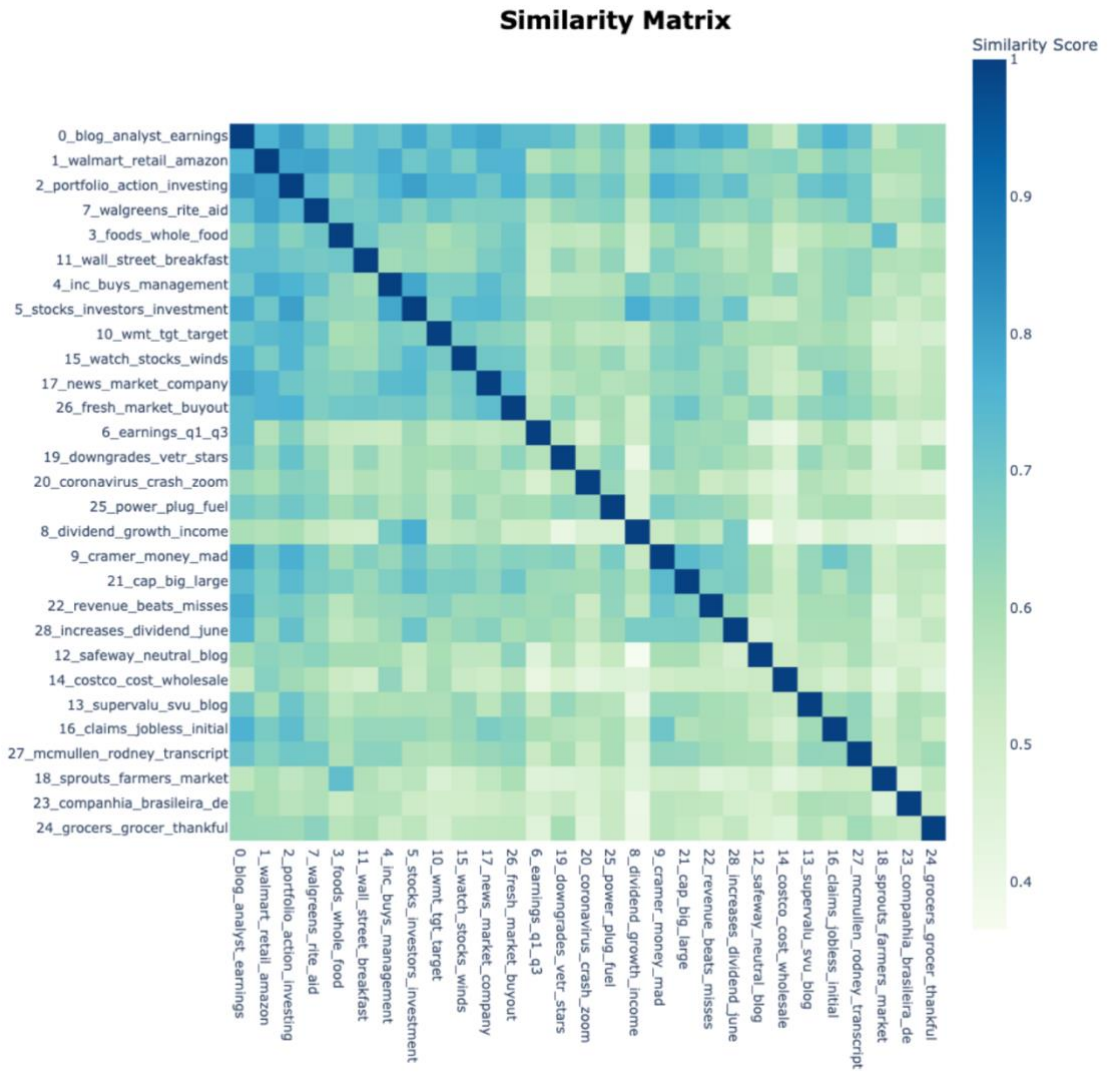


Figure 11-Similarity Matrix

After topic reduction, interpretability increases, darker colours represent higher similarity between topics, for instance, topic 7 and 1 (about competitors walgreens and Walmart).

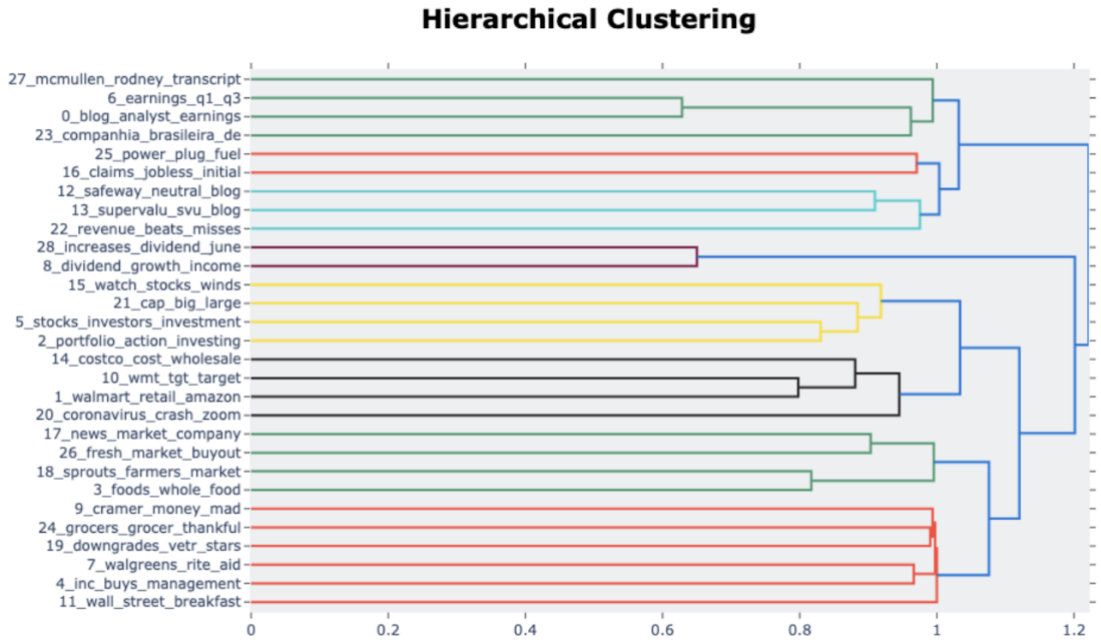


Figure 12-Hierarchical clustering

Figure 12 shows topics clustered together, like competitors Costco,Amazon in 14,1, financial performance in topics 8,28.

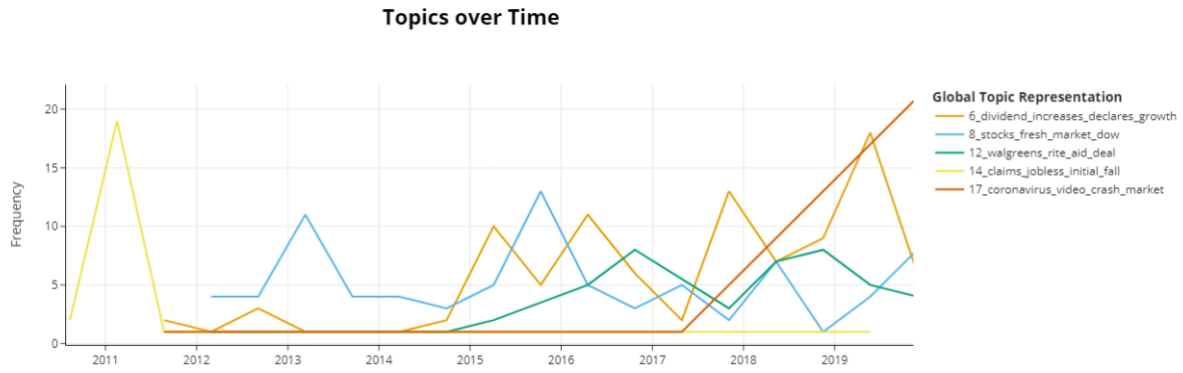


Figure 13-Topics over time

Figure 13 shows how selected topics changed over the years. As expected, coronavirus is mentioned from 2018. Dividend growth topics fluctuates throughout the years along with stocks.

Conclusion

In conclusion, news articles and disclosures could contain important forward-looking information for stakeholders' decision making instead of reliance on irrelevant historical financial information, extracting useful topics. Using semi-supervised learning, I inferred the stock price movement through training and evaluated its performance in another time period, although model performance is poor.

Limitations:

Future papers should focus on extending the implementation to other data sources such as Twitter tweets, Reddit posts, RSS feeds and Telegram channels focused at financial news and take stratified sample from different stocks for topic modelling (see Appendix).

The time consuming labelling of stock symbol for each document in dataset and susceptibility to mistakes would garner:

- Obtaining from sources only related to specific stock
- Use Name Entity Recognition
    - Self-supervised labelling (comparing between headline and text document), starts with upper-case character
    - Pre-trained models

Although computationally more expensive, using batches on larger tokens of headlines may generate better results, while a robust evaluation, hyper-parameter tuning is better.

Appendix

In future, I would take documents (headlines) from each of the stocks with more than 1000 documents (considered as representative) and stratify sample - taking randomly (with deterministic random state=0) 10 documents each, and apply topic modelling on it, using this method:

```python
df_headlines_mt_1000 = df.groupby("Stock").filter(lambda x: len(x) > 1000)
```

Python

```
# df_sampled = df.groupby('Stock', group_keys=False).apply(lambda x: x.sample(random_state = random_state, frac = 0.05))
df_sampled = df_headlines_mt_1000.groupby('Stock', group_keys=False).apply(lambda x: x.sample(random_state = random_state, n=10))    "random_state" is not de
```

References

[1] A. Udupa, K. N. Adarsh, A. Aravinda, N. H. Godihal, and N. Kayarvizhy, 'An Exploratory Analysis of GSDMM and BERTopic on Short Text Topic Modelling', in *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP)*, Dec. 2022, pp. 1–9. doi: 10.1109/CCIP57447.2022.10058687

[2] 'SN Top 10: UNFI, Kroger, SpartanNash top the week's headlines', *Supermarket News*, Jul. 18, 2022. Available: https://www.supermarketnews.com/news/sn-top-10-unfi-kroger-spartannash-top-weeks-headlines. [Accessed: Jul. 22, 2023]

[3] 'Daily Financial News for 6000+ Stocks'. Available: https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests. [Accessed: Jul. 22, 2023]

[4] Yahoo!, 'Market data and research tools available in Yahoo Finance'. Available: http://help.yahoo.com/kb/SLN24381.html. [Accessed: Jul. 22, 2023]

[5] R. Egger and J. Yu, 'A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts', *Front. Sociol.*, vol. 7, 2022, doi: 10.3389/fsoc.2022.886498. Available: https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498. [Accessed: Jul. 22, 2023]

[6] M. P. Grootendorst, 'Quick Start - BERTopic'. Available: https://maartengr.github.io/BERTopic/getting_started/quickstart/quickstart.html. [Accessed: Jul. 22, 2023]

[7] A. Abuzayed and H. Al-Khalifa, 'BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique', *Procedia Comput. Sci.*, vol. 189, pp. 191–194, 2021, doi: 10.1016/j.procs.2021.05.096

[8] M. J. Sánchez-Franco and M. Rey-Moreno, 'Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings', *Psychol. Mark.*, vol. 39, no. 2, pp. 441–459, 2022, doi: 10.1002/mar.21608

[9] Priyanka, 'Understanding BERTopic Intuitively', *Medium*, Feb. 01, 2023. Available: https://levelup.gitconnected.com/understanding-bertopic-intuitively-b3787f104bda. [Accessed: Jul. 22, 2023]

[10] S. Prabhakaran, ‘Lemmatization Approaches with Examples in Python’, *Machine Learning Plus*, Oct. 02, 2018. Available: https://www.machinelearningplus.com/nlp/lemmatization-examples-python/. [Accessed: Jul. 22, 2023]

---

[i] Maximal Marginal Relevance
[ii] Uniform Manifold Approximation and Projection
[iii] Corporate social responsibility