

# Characterizing Financial Market Coverage using Artificial Intelligence

Jean Marie Tshimula,<sup>\*1,2</sup> D’Jeff K. Nkashama,<sup>\*1</sup> Patrick Owusu,<sup>\*1,3</sup> Marc Frappier,<sup>1</sup> Pierre-Martin Tardif,<sup>1</sup> Froduald Kabanza,<sup>1</sup> Armelle Brun,<sup>3</sup> Jean-Marc Patenaude,<sup>4</sup> Shengrui Wang,<sup>1</sup> Belkacem Chikhaoui<sup>2</sup>

<sup>1</sup>Department of Computer Science, Université de Sherbrooke, QC J1K 2R1, Canada

<sup>2</sup>LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada

<sup>3</sup>LORIA, Université de Lorraine, 54000 Nancy, France

<sup>4</sup>Laplace Insights, QC J1H 1P9, Canada

shengrui.wang@usherbrooke.ca

## Abstract

This paper scrutinizes a database of over 4900 YouTube videos to characterize financial market coverage. Financial market coverage generates a large number of videos. Therefore, watching these videos to derive actionable insights could be challenging and complex. In this paper, we leverage Whisper, a speech-to-text model from OpenAI, to generate a text corpus of market coverage videos from Bloomberg and Yahoo Finance. We employ natural language processing to extract insights regarding language use from the market coverage. Moreover, we examine the prominent presence of trending topics and their evolution over time, and the impacts that some individuals and organizations have on the financial market. Our characterization highlights the dynamics of the financial market coverage and provides valuable insights reflecting broad discussions regarding recent financial events and the world economy.

## 1 Introduction

Financial markets, especially the stock market, enjoy substantial coverage day-to-day on digital platforms such as YouTube. Besides the presenters, experts with much understanding of the stock markets work as contributors or panelists who share their perspectives on various topics. On YouTube, channels such as Yahoo Finance’s Stock Market Coverage provide a wealth of information about the development of financial market events, that can allow the audience to get informed on trending topics, among others. Most studies on the impact of stock news, for instance, on stock prices focused on using headlines from renowned news agencies or blog posts (Jariwala et al., 2020; Velay and Daniel, 2018; Nemes and Kiss, 2021). The news coverage particularly gives a topic context and meaning (Chipidza et al., 2022), much as financial television

(TV) programs such as CNBC Markets. However, in comparison to traditional news coverage, the dissemination of financial and economic news is either a segmented section or a dedicated channel on TV.

While most financial market studies focused on data sources such as news headlines, financial reports, and social media posts, continued news coverage of any kind have had limited usage for the purpose of analysis. In particular, one may argue the authenticity of social media posts from either Facebook or Twitter, as misinformation is ubiquitous on such platforms (Kogan et al., 2021). With media coverage, not only is it factually oriented to decrypt market news and events, an advantage is that the information is fact-checked. Besides this, the viewpoints of renowned experts can get contradicted, backed, or completed by journalists or other high-profile persons in the world of finance and economics.

Given the popularity of YouTube, the publicly available videos constitute a reliable source of data for further analysis. However, the challenge of analyzing videos, in general, requires either capturing image frames or manipulating snippets of an entire video (Snelson et al., 2021). The social science field is an example where videos are a data source for analysis; these are either non-verbal or otherwise (Luff and Heath, 2012).

In this paper, we build corpora of transcribed videos on YouTube that focus on the financial market and the economy. We used OpenAI’s Whisper (Radford et al., 2022) to transcribe videos to texts since market coverage generates a large volume of video data. Consequently, watching tons of financial market coverage videos to derive actionable insights can be challenging and complex. To this end, we transcribe market coverage videos to texts for simplifying analyses. Specifically, we use a topic modeling approach for generating topics related to the markets. Further, we perform an n-gram

\*These authors contributed equally to this work

analysis to understand the coverage narratives and extract the most frequently mentioned persons and organizations in the market coverage using named entity recognition. It should be noted that we kept topics related to the economy and markets.

**Background and Related Works.** The characterization of financial and economic news has been explored in numerous aspects, including emotions and sentiments (Schumaker et al., 2012; Griffith et al., 2020) from social media posts to news headlines (Mitra and Mitra, 2011; Bukovina, 2016). The role of news has been covered in (Baker et al., 2021) where the authors pose the question “*What drives big moves in national stock markets?*”. According to the study, news about US economic and policy developments significantly impacts worldwide equity markets. Considering previous works, most have centered on how media coverage affects financial markets (Fang and Peress, 2009; Dougal et al., 2012; Strycharz et al., 2018). In comparison with previous works on the effects of media coverage on financial markets, we focus on how financial and economic news coverage narrations between multiple platforms are similar. Our work is closely related to the studies of (Piao, 2015; McBeth et al., 2018; Bhargava et al., 2022), who compare the narratives of news coverage. Relative to these works, we investigate the evolution of the popular topics addressed over time by financial media channels on YouTube and discover similarities between the topics addressed by different media channels. Additionally, we show that financial news coverage is centered on organizations and individuals: where individuals are either heads of organizations or knowledgeable panelists or experts in finance and economics. We focus on the following two research questions. *RQ1*: How are major financial events identified through language use within news coverage topics? *RQ2*: To what extent do news coverage topics exhibit content coordination regarding major financial events and entities (such as organizations and individuals) across different news channels?

Specifically, this work makes the following contributions. (1) We show how effective our data collection and pre-processing strategy is for gathering digital videos and generating text information, which relates to the financial market and economic discourse. (2) We compare the narratives between reliable media channels; one of the advantages of utilizing datasets from these media channels is that

Table 1: Data summary of the collected coverage

Media	Tot. collected files	Total hours	Avg. time per file
BLW	744	171.16	14 minutes
BSM	3885	398.15	6 minutes
YFM	318	2467	8 hours

they do not stem from bots. (3) We investigate the evolution of the topics addressed over time and examine the most frequently mentioned entities (organizations and persons) in financial market coverage and discover similarities between topics. (4) We publicly release our code as open source to support continued development.<sup>1</sup>

## 2 Methods

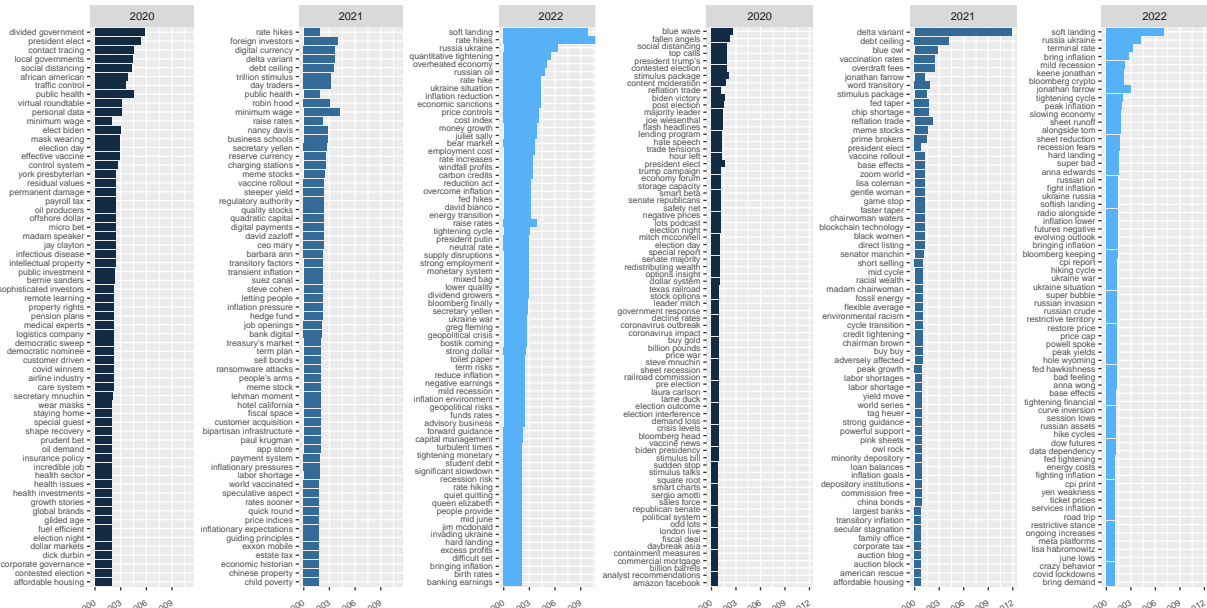
### 2.1 Data collection

The data we used in the study was collected from the YouTube channels of Yahoo Finance and Bloomberg Markets and Finance and transcribed using the OpenAI’s Whisper speech-to-text model described in (Radford et al., 2022). Note that our choice of using Yahoo Finance and Bloomberg is motivated by the fact that they (i) are among the world leaders in business news and real-time financial market coverage, (ii) provide financial news, data, and commentary including stock quotes, press releases, financial reports, original content, and video to the world of finance every Monday–Friday from 9 am to 5 pm (ET), and (iii) host world-class specialists to express their opinions and discuss market news and events; additionally, they are freely accessible on YouTube. They decompose the markets and real-life financial issues for individual investors, industry leaders, and those seeking to invest in their future.

Specifically, we collected one of the YouTube playlists of the Yahoo Finance Market’s official channel called Stock Market Coverage, from 02 January 2020 to 30 September 2022. We extracted two different YouTube playlists from Bloomberg and Finance’s official channel, namely Wall Street Week, and Stock Market News and Analysis, for the same time period.

**Yahoo Finance Stock Market Coverage (YFM)** receives top names in finance and economics to scrutinize the latest market news and contribute with cogent evidence to explain the development of the market events, identify any untapped needs in the marketplace, and provide ad-

<sup>1</sup>[https://github.com/djeffkanda/market\\_coverage\\_analysis](https://github.com/djeffkanda/market_coverage_analysis)



(a) Bloomberg Wall Street Week (BLW)

(b) Bloomberg Stock Market News and Analysis (BSM)

Figure 1: Bi-grams with the highest tf-idf from Bloomberg data. Note that  $x$ -axis represents tf-idf values.

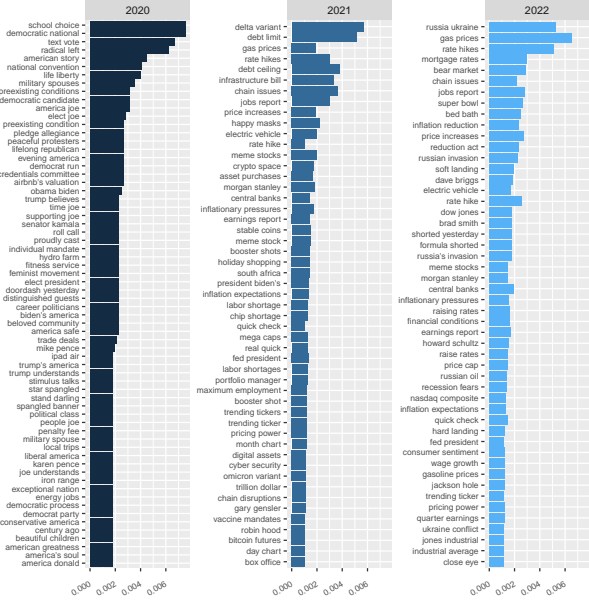


Figure 2: Bi-grams with the highest tf-idf from YFM. Note that  $x$ -axis represents tf-idf values.

vanced analyses and opinions.

**Bloomberg Wall Street Week (BLW)** hosts influential personalities in finance and economics to talk about the week’s biggest issues on Wall Street.

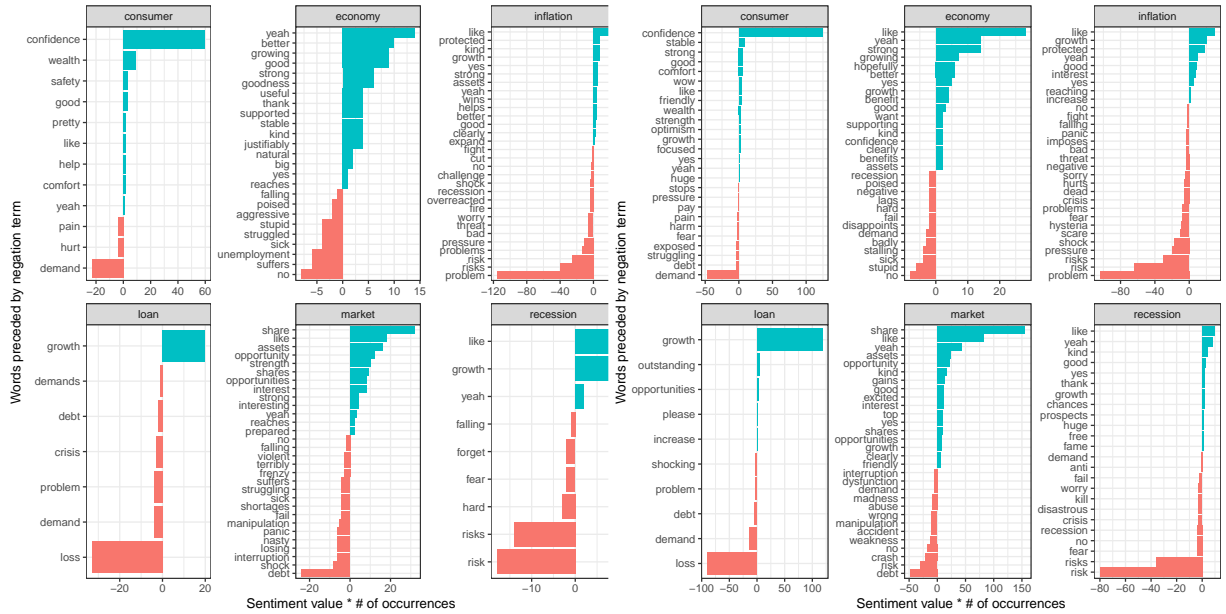
**Bloomberg Stock Market News and Analysis (BSM)** is a playlist of Bloomberg Markets and Finance’s YouTube official channel where experts discuss the latest market news and effectuate market analysis in real-time coverage.

Table 1 summarizes statistics of the collected market coverage. For the three targeted YouTube

playlists, BLW, BSM, and YFM, respectively, we extracted 744, 3885, and 318 videos for which the total hours approximate 171.16, 398.15, and 2467 and the average duration of videos counts 14 minutes, 6 minutes and 8 hours.

We utilized the OpenAI’s Whisper, a speech recognition model, to transcribe audios of the collected data to text corpora (Radford et al., 2022). Speech recognition remains a challenging problem in artificial intelligence and machine learning (Chiu et al., 2018; Qin et al., 2019; Zhang et al., 2020). In a step toward solving it, OpenAI introduced Whisper, an automatic speech recognition system that approaches human-level robustness and accuracy in English speech recognition. Whisper outperformed the state-of-the-art speech recognition systems by leaps and bounds and has received immense interest for its multilingual transcription and translation capabilities spanning nearly 100 languages. Whisper was trained on 680,000 hours of multilingual and ‘multitask’ data collected from the web, which lead to improved recognition of unique accents, background noise, and technical jargon. One of the advantages of Whisper is that it performs well even on diverse accents and technical language and is almost human-level in terms of recognizing speech even in extremely noisy situations. The architecture and the performance of Whisper over other speech recognition systems are briefly explained in its original paper (Radford et al., 2022).

Specifically, we utilized the transcribed corpora



(a) Bloomberg Wall Street Week (BLW)

(b) Bloomberg Stock Market News and Analysis (BSM)

Figure 3: Words preceded by either *consumer*, *economy*, *inflation*, *loan*, *market* or *recession* that had the greatest contribution to sentiment values, in a positive or negative direction in Bloomberg coverage.

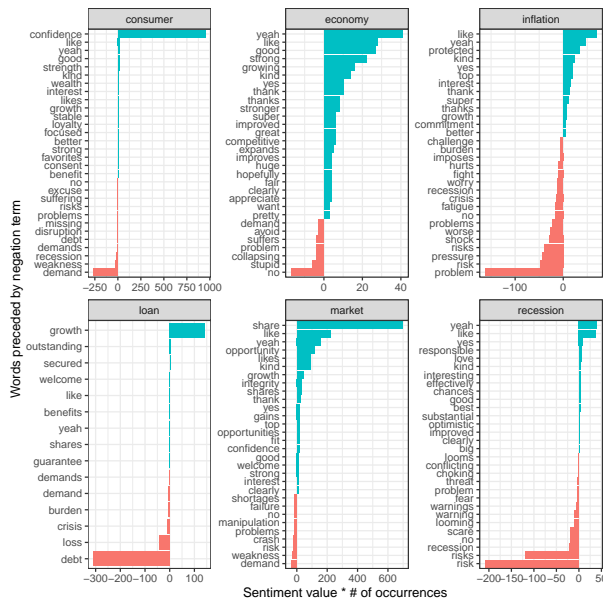


Figure 4: Words preceded by either *consumer*, *economy*, *inflation*, *loan*, *market* or *recession* that had the greatest contribution to sentiment values, in a positive or negative direction in YFM.

(Table 1) to extract insights using natural language processing techniques including n-gram analysis (§2.2), topic modeling (§2.3) and named entity recognition (§2.4). We removed stopwords, numbers and special characters for performing n-gram analysis §2.2 and topic modeling §2.3.

## 2.2 N-gram analysis

We analyzed n-grams to extract important insights in text transcriptions to understand language use within news coverage narratives. We extracted bi-grams from text transcriptions of financial market coverage by leveraging the vectors based on the term frequency-inverse document frequency (*tf-idf*) technique (Ramos et al., 2003; Gebre et al., 2013). Specifically, we utilized *tf-idf* as a statistical measure to evaluate how important a word is to each text transcription in the corpus; we converted each text transcription into its bag-of-words representation and computed the *tf-idf* value of each word using the standard formula,  $tf-idf = (1 + \log n_{w,t}) \times \log \frac{T}{T_w}$ , where the *tf-idf* value of word *w* in text transcription *t* is the log normalization of the number of times the word occurs in the text transcription ( $n_{w,t}$ ) times the inverse log of the number of text transcriptions *T* and  $T_w$  the number of text transcriptions containing word *w*.

## 2.3 Topic modeling

Topic modeling refers to the machine learning task of automatically discovering the abstract ‘topics’ that occur in a collection of documents, and one popular topic modeling technique is known as latent Dirichlet allocation (LDA) (Blei et al., 2003). LDA is a probabilistic model that identifies latent topics in a document and can be trained using collapsed Gibbs sampling. Fundamentally, LDA as-

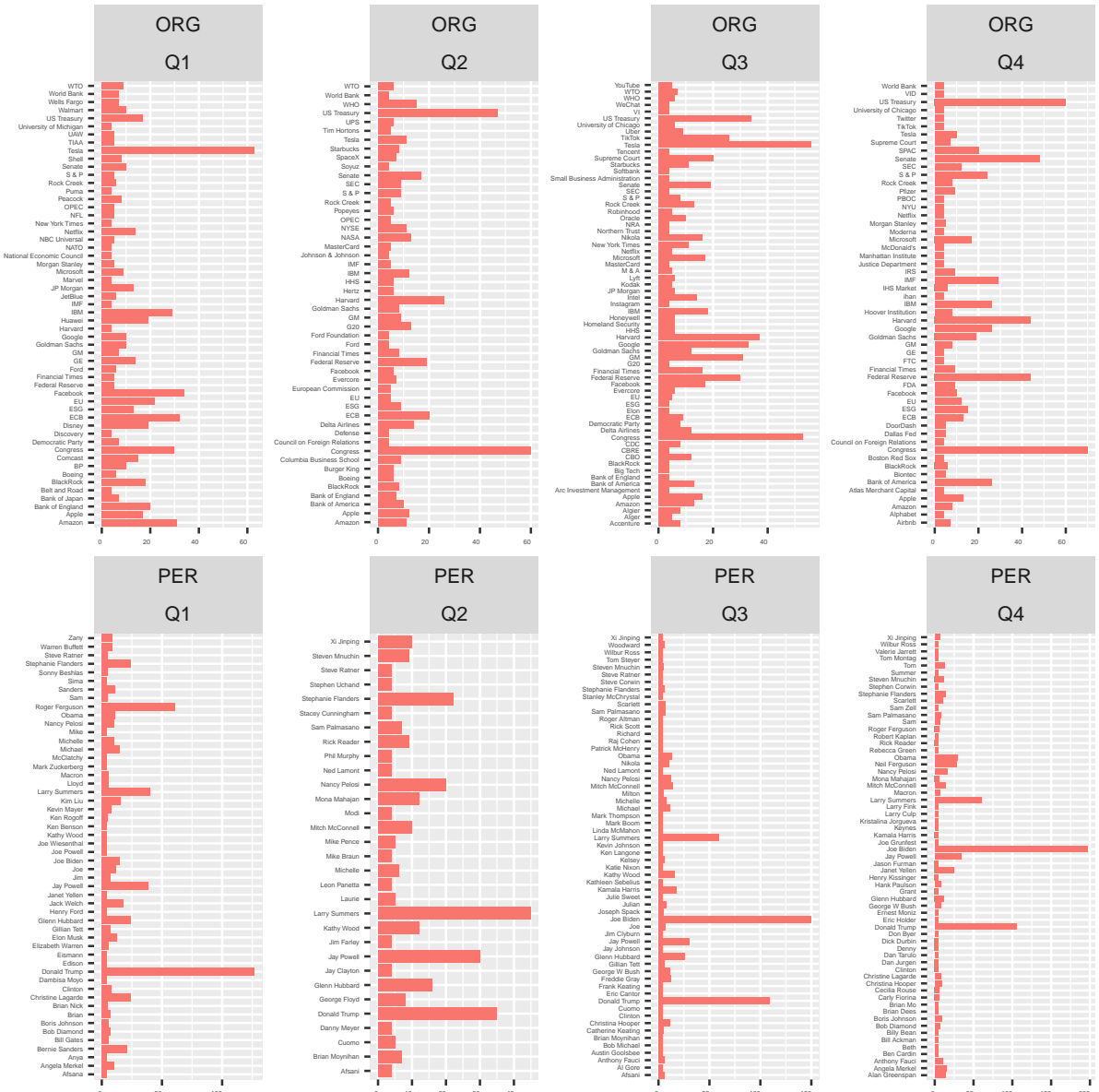


Figure 5: NER results of BLW for 2020

sumes  $k$  underlying topics, each of which is a distribution over a fixed vocabulary. While LDA models topics from text corpora (Angelov, 2020), it basically suffers from several shortcomings, including difficulty in setting the parameter  $k$  (which refers to the number of topics to produce semantically meaningful results), a deficiency in handling short texts (Banda et al., 2021), in capturing the contextual meaning of sentences (Žuk and Žuk, 2020), and its inability to model topic correlations and the evolution of topics over time (Wang and McCallum, 2006).

To overcome these limitations, the new generation of topic models (Peinelt et al., 2020; Bianchi et al., 2020; Angelov, 2020; Grootendorst, 2022)

utilize pre-trained representations such as BERT to enable topic modeling (i) to consider the contextual meaning of sentences for supporting the results in order to match the adequate topics and (ii) to include more features for efficiently modeling topic correlations and topic evolution over time. Recent pre-trained contextualized representations like BERT have pushed the state-of-the-art in several areas of natural language processing due to their ability to expressively represent complex semantic relationships from being trained on massive datasets. BERT is a bidirectional Transformer-based pre-trained contextual representation using masked language modeling objective and next sentence prediction tasks (Devlin et al., 2018). The

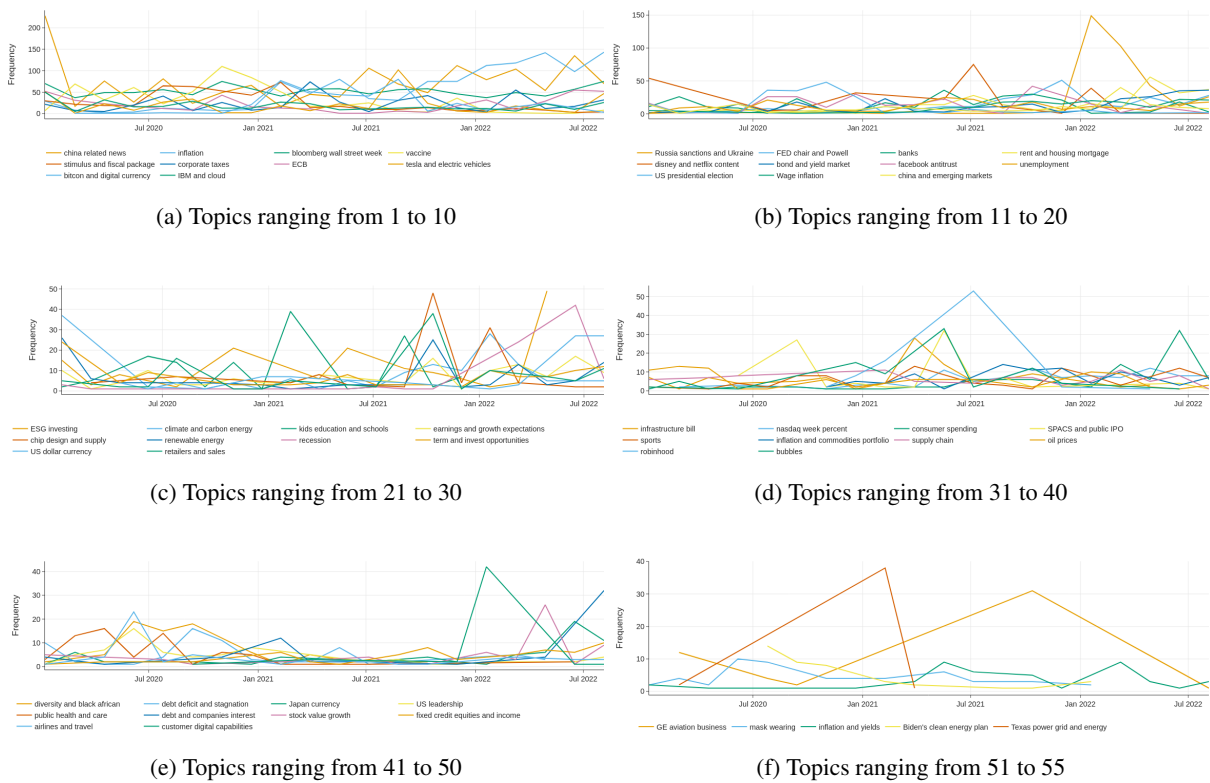


Figure 6: Top 55 topics with frequencies over time extracted from Bloomberg Wall Street Week

significant advantage of BERT is that it simultaneously gains the context of words from both the left and right context in all layers. To this end, BERT utilizes a multi-layer bidirectional Transformer encoder, where each layer contains multiple attention heads.

In this paper we use BERTopic (Grootendorst, 2022) to generate topics addressed in financial market coverage, analyze the evolution of these topics over time and discover similarities between the topics addressed in BLW, BSM and YFM (*RQ1*). BERTopic leverages BERT embeddings and a class-based term frequency-inverse document frequency to create dense clusters to detect unique topics. In addition, BERTopic generates the topic representations at each timestamp for each topic. The traditional LDA model requires a predefined  $k$  (the number of topics) for algorithms to cluster corpus around  $k$  topics (Blei et al., 2003). BERTopic does not require a predefined  $k$ , reducing the need for various iterations of model finetuning. The performance of BERTopic over LDA-like models and other topic modeling techniques is reported in (Grootendorst, 2022).<sup>2</sup>

<sup>2</sup>The Python package of BERTopic: <https://github.com/MaartenGr/BERTopic>

## 2.4 Named entity recognition

Named entity recognition (NER) aims at finding and categorizing specific entities in text with their corresponding semantic types such as person names, organizations (such as companies, government organizations, etc.), locations (such as cities, countries, etc.), or date and time expressions (Li et al., 2020; Perera et al., 2020). In this paper, we utilized NER to extract the names of persons and organizations mentioned in financial market coverage, and map the frequency of the most mentioned entities over time (*RQ2*). The rationale behind the extraction of NER entities is to identify entities that constitute the center of attention in the financial market and dominate the financial world at a specific time-step. This could support the understanding of the evolution of the topics addressed over time and indicate the entities around which the topics are concentrated in. Note that we removed some names of Bloomberg and Yahoo anchors that appeared in the NER results.

We used a fine-tuned BERT model called bert-base-NER.<sup>3</sup> In this paper we used the initial model of bert-case-NER without modifying its architec-

<sup>3</sup>Find the official page of bert-base-NER on HuggingFace, [dslim/bert-base-NER](https://huggingface.co/dslim/bert-base-NER)

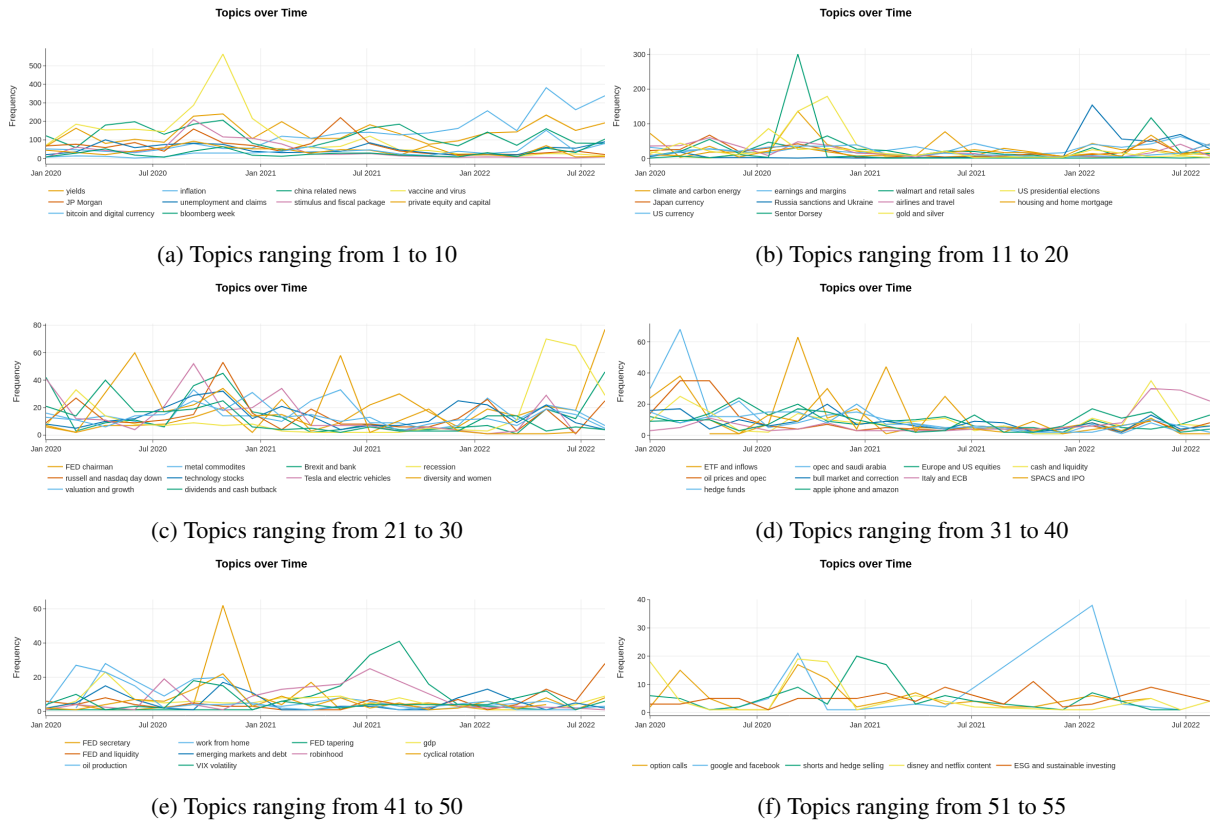


Figure 7: Top 55 topics with frequencies over time extracted from Bloomberg Stock Market News and Analysis

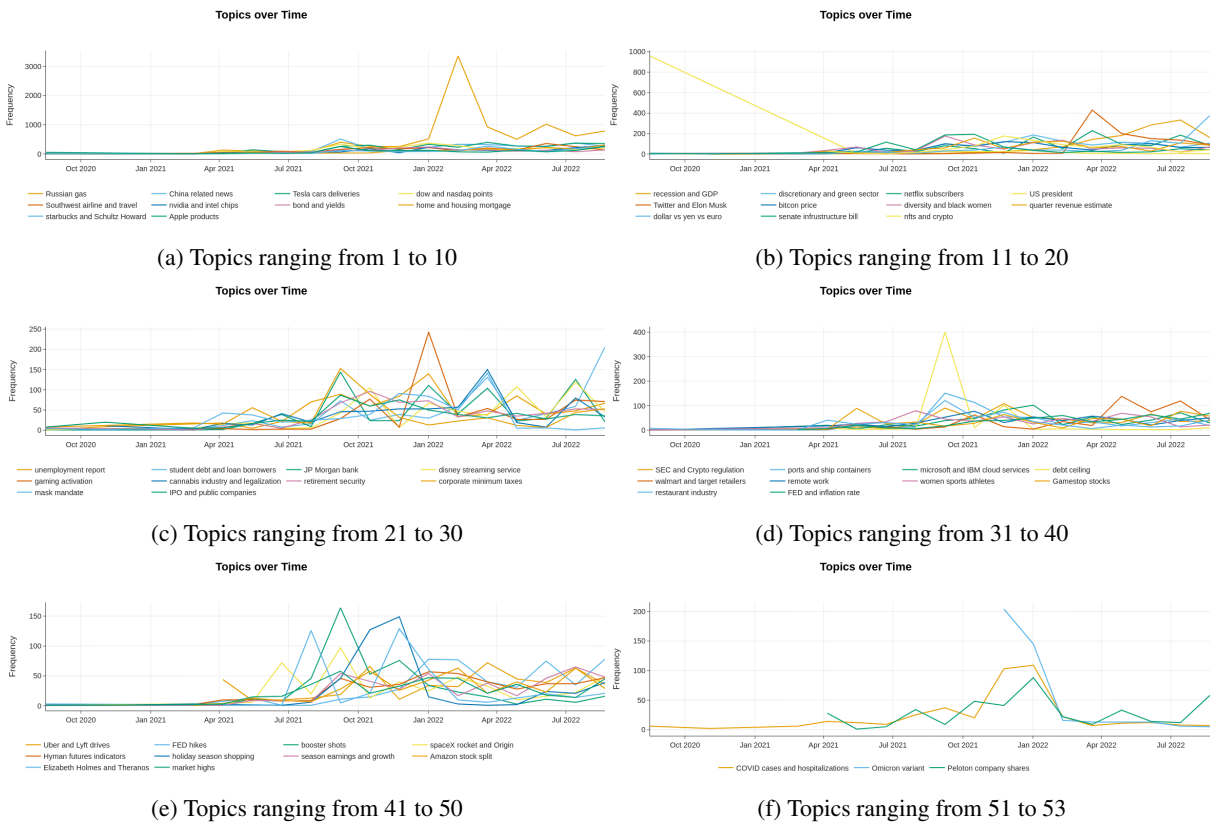


Figure 8: Top 55 topics with frequencies over time extracted from Yahoo Finance Stock Market Coverage

ture or implementation. It is important to note that bert-base-NER is ready to use for NER and achieves state-of-the-art performance for the NER task, and it has been trained to extract four types of entities: location (LOC), organizations (ORG), person (PER) and miscellaneous (MISC). Specifically, bert-base-NER is a version of the bert-base-cased model that was fine-tuned on the English version of the standard CoNLL-2003 Named Entity Recognition dataset (Sang and De Meulder, 2003; Devlin et al., 2018).

### 3 Results

This section first investigates the relationships between words using n-grams (n=2) and language use within news coverage narratives and provides sentiments of n-grams indicating economic and financial concerns. Second, topic modeling and name entity recognition are performed to examine the evolution of topics over time and discern the coverage narratives, and then understand the implication of the most mentioned persons and organizations in the stock market.

**Bi-gram analysis.** As mentioned in §2.2, we sort to determine the link between words by looking at which words typically come after others right away. By utilizing the *tf-idf* statistical measure, we identify the importance of each consecutive sequence of words in each year’s corpus. Figures 1a, 1b and 2 show the top 55 bi-grams drawn from the market coverage. Generally, each year, we observe the bi-grams are primarily on economic and financial markets-related topics as well as some pertaining issues that happened during the years.

While examining the results obtained from the BLW, BSM and YFM datasets, we observe that the majority of bi-grams indicate economics-related topics except for “*divided government*”, “*delta variant*”, “*democratic national*” and “*blue wave*”, among others. Particularly, we note that these datasets assimilate topics/events pertaining to aspects of finance that have major impacts on the economy, for example, prices. Besides these, other interesting financial discourses are centered on the “*game stop*” fiasco in 2021 (Malz, 2021); we observed this as a bi-gram in BSM. Interestingly, we note the presence of the bi-gram “*digital assets*” along with “*cyber security*”. Recent turmoil in the cryptocurrency market has underlined the critical risks involved with investing in or engaging with digital assets. Digital assets raise cybersecurity

concerns requiring regulatory controls and measures to protect individuals from cybercrime and other critical risks (Chaisse and Bauer, 2018). We observe an important number of cryptocurrency-related results, including bi-grams “*bitcoin futures*” and “*crypto space*” in 2021, topics “*bitcoin price*” and “*nfts and crypto*” in Figure 8 and the topic “*bitcoin and digital currency*” in Figures 6 and 7.

We report bi-grams highlighting recent events occurring in Ukraine as well as their continuous in the early months; these bi-grams include “*russia ukraine*”, “*russia oil*” and “*ukraine conflict*” are especially inline with the commodities market.<sup>4</sup> Additionally, we note that the coverage bi-grams also identify persons and organizations including “*central banks*”, “*paul krugman*” and “*gary gensler*”. For instance, Paul Krugman is an economist and a contributor on Bloomberg<sup>5</sup> and Gary Gensler Chairperson of the U.S. Securities and Exchange Commission since 2021.<sup>6</sup> Overall, we find that some bi-grams depict the language use in coverage which attributes to events such as “*president elect*” or “*rate hikes*”; each referring to the general elections and (imminent) announcement of interest rate hikes and discussions on these type of events.<sup>7</sup>

Figure 3a, 3b and 4 show the top 6 economic (financial)-related keywords identified in the bi-gram that best describes the financial markets. The keywords (*consumer*, *economy*, *inflation*, *loan*, *market* and *recession*) are not exhaustive and can be expanded. The choice of these keywords is arbitrary, we believe that they reflect and pertain to broad discussions regarding recent events. We examine how frequently sentiment-associated words are preceded by these keywords, which attribute to positive or negative sentiments; with positive or negative values indicating the direction of the sentiment. We note that bi-grams stemming from the previously mentioned keywords identify the most common economic events. For instance, the bi-gram of “*demand consumer*” has a negative sentiment while “*confidence consumer*” has a positive sentiment. These bi-grams reflect the events of supply-chain issues or consumers’ ability to buy items.<sup>8</sup> Further, “*inflation*” and “*economy*” discourse cite either “*growth/growing/good*” or “*risk(s)/worse/stalling*” painting the picture of posi-

<sup>4</sup>[shorturl.at/lqWZ3](https://shorturl.at/lqWZ3) Accessed 23 December 2022

<sup>5</sup>[shorturl.at/dHIX8](https://shorturl.at/dHIX8) Accessed 23 December 2022

<sup>6</sup><https://www.sec.gov/>

<sup>7</sup>[shorturl.at/mwBGL](https://shorturl.at/mwBGL) Accessed 23 December 2022

<sup>8</sup>[shorturl.at/brJOP](https://shorturl.at/brJOP) Accessed 23 December 2022



tive and negative sentiments, respectively. The bi-gram of “*stalling economy*” essentially describes one with a growth rate below some threshold level. Thus, the possible effects of the COVID-19 pandemic. The bi-gram of “*growth loan*” is the maximum positive sentiment across BLW, BSM and YFM data. Our analysis indicates the right direction of such a bi-gram. However, the financial keyword, “*market’s*” bi-gram has “*share*” as the most positive sentiment and with “*demand/debt*” as the opposite. A discussion relating to market share could be attributed to an organization as the bi-grams of Figures 1a, 1b and 2 identify some organizations. Interestingly, we observe that the NER analysis also identified such organizations.

**Named entity recognition.** Within the context of (financial) news coverage, individuals (or persons) are either introduced as panelists or as a contributor or mentioned (cited) to affirm a statement. Likewise, some individuals are associated with some organization (or institution), or sometimes discussions are centered around an organization based on what might be trending. To distinguish between persons and organizations from our corpora, we employed a NER model as described in §2.4. Figure 5 shows the distinct entities for the BLW corpus in each quarter of 2020.

Each quarter shows the frequencies of the top 60 entities (organizations and persons). A closer look at all the quarters’ organizations reveals the following organizations having the highest mentions or often discussed: *Tesla*, *Congress* and *US Treasury*. Note that *Congress* represents both the “House of Representatives” and “Senate”. Within the various quarters, we noticed that some major technology companies were frequently in discussions: *IBM*, *Huawei*, *Facebook*, *Apple* and *Amazon*. The *Congress*, for example, were often concerning stimulus package discussions.<sup>9</sup> We further observed that the identified organizational entities are either governmental or private financial institutions, such as *US Treasury*, *Goldman Sachs* and *Rock Creek*. Besides, there were organizations of global and continental significance during the coverage. For example, the *World Bank*, *OPEC* and *ECB*—the European Central Bank.

Similarly, the news coverage on persons ranges from world leaders or politicians to investment moguls to financial experts to heads of institutions

and others. In the first quarter of 2020, we noticed that the name of *Donald Trump*, the then president of the United States of America (USA), was frequently mentioned; this suggests that the events of January 6, 2020, and subsequent events had tremendous discussions in the financial and economic news space. We also noticed an important frequency around the name of *Roger Ferguson*, the former president and Chief Executive Officer (CEO) of *TIAA*—organization and a contributor on BLW. In the subsequent quarters, *Larry Summers*, a renowned financial expert and contributor; *Joe Biden*, the current president of the USA; and *Donald Trump* were highly mentioned. Concerning *Larry Summers*, he provides insight into how prospective the economic and financial outlook would be based on some announcements. Of particular notice was *George Floyd* in the financial news coverage in the second quarter of 2020; his murder sparked numerous protests and moments of reckoning that reverberated far beyond the United States. Based on the NER sample result, we observed that financial news coverage does not only cover finance and economic topics but also general topics. In the next section, we identify some common topics from the news coverage.

**Topic modeling.** Figures 6, 7 and 8 show the top 55 salient topics from the BLW, BSM and YFM, respectively. Figures are organized into six sub-figures at the rate of ten topics per sub-figure to provide a better visualization of the frequency of topics over time for the period from January 2020 to September 2022.

Figure 6a shows the ten most topics addressed in BLW. These topics include “*inflation*”, “*vaccine*”, “*China-related news*” and “*Tesla and electric vehicles*”. Particularly, for the topic “*Tesla and electric vehicles*”, a high spike was observed in early 2020, followed by a drop in frequency over the first quarter of 2021. Even though electric vehicles seemed less frequently discussed in favor of topics related to the vaccine and COVID-19, we observed that the discussions around electric vehicles remained one of the most salient topics in the market coverage. We noticed many spikes in Figure 6a and Figure 7a for the topic “*vaccine*” during 2020 and the topic “*mask-wearing*” in Figure 8c. One of the reasons that could partly justify this observation is that pharmaceutical companies such as Pfizer and BioNTech started research on developing vaccines for COVID-19 during that period and announced

<sup>9</sup>[www.bloomberg.com/news/articles/2020-03-25/what-s-in-congress-2-trillion-coronavirus-stimulus-package](https://www.bloomberg.com/news/articles/2020-03-25/what-s-in-congress-2-trillion-coronavirus-stimulus-package)

promising results. In November 2020, Pfizer announced the vaccine releases, followed by a vaccination campaign worldwide.<sup>10</sup> The COVID-19 pandemic has caused major social and economic impacts on the lives of people across the world. One of the direct impacts includes unemployment in the labor market (Figure 6b, 7a, and 8c), and inflation, in the financial market. Figure 8c shows an increase in frequency over time for the topic “inflation” from 2021 to 2022. We note that this topic received considerable attention in the market coverage along with its related topics, such as “home and housing mortgage”, due to the increase in mortgage rates, and “recession”.

Figures 6b, 7a and 8c highlight the evolution of topics over time for topics such as “Disney and Netflix content” and “Russia sanctions and Ukraine”. The Russian invasion of Ukraine in early 2022 caused knock-on effects worldwide. Sanctions imposed on Russia by the United States and other countries engendered multilateral effects on the global economy in general and the stock market in particular. This reason could be retained as a compelling justification to partly explain the many high spikes that we observed for the topic of Russia and Ukraine. The topic “Disney and Netflix content” indicates its large surge during the period of the first COVID-19 lockdown. Note that lockdown was one of the restrictive measures taken by governments to contain the ongoing pandemic. During the lockdown period, many people spent most of their time on streaming platforms as they could not go out. Online streaming platform subscriptions have increased along with their corresponding stock price.

## 4 Discussion

News coverage provides contest and analysis needed to aid viewers in ascertaining further insight lacking from other news sources (newspapers or blogs) through anchors and guests (experts’) discussions. In this paper, we collected news coverage data from YouTube and Bloomberg related to financial and economic news to identify the most discussed topics from transcribed video news coverage.

The primary goal of our research was to identify the similarities of news coverage topics regarding major financial events across different news channels. Our findings describe the usefulness of con-

sidering video (visual) as a data source. By analyzing the similarity between other channels, we observe some related bi-gram keywords and entities (organizations and persons). The bi-gram provides an overview of the structure of language use during news coverage through discussions and headlines briefing of news segments. Our results find that news coverage evolved, and discussions were often centered more on recent events surrounding specific financial markets.

Secondly, we identified major financial events through the evolution of topics over time and their frequencies. Our topic models broadly reflect the evolution and variation of topics related to financial events. Important to note are the global events documented in various studies that are in tandem with financial markets, such as the Russo-Ukrainian War (Lo et al., 2022). Prior work found the effect of news coverage on trading and prices (Engelberg and Parsons, 2011; Haroon and Rizvi, 2020), while our results identify the narrative of news coverage without any relation to either trading behavior or price volatility. Our results can be used to create dashboards portraying outputs stemming from financial market coverage from various reliable media channels. This can help anticipate “investment” actions or predict market pricing based on the news coverage and identify the most frequently cited entities to make a good investment choice. Further, the results investigated the less frequently cited entities, which one can keep a constant eye on or keep on track to ensure if they constitute a new market opportunity and if something might skyrocket overnight.

## 5 Conclusion

In this paper, we characterize financial market coverage from YouTube. To this end, we utilize OpenAI’s Whisper speech-to-text model to generate a text corpus of market coverage YouTube videos from Bloomberg and Yahoo Finance. Then, we use natural language processing to gain insights into language usage in financial market coverage. Additionally, we investigate the prevalence and evolution of trending topics and the influence of certain persons and organizations on the financial market. We discover similarities between topics and exhibit content coordination regarding major financial events. Through this characterization, we gain a better understanding of the dynamics of financial market coverage and valuable insights into current

<sup>10</sup>Pfizer and BioNTech Announce Vaccine Candidate Against COVID-19. [shorturl1.at/mHNV4](https://shorturl1.at/mHNV4)

financial events and the global economy. We show how our findings can be used to predict market performance and pricing and to support investment actions and decision-making. In the future, we would like to experiment with market forecasts using a holistic model that combines financial market coverage and stock prices and includes features such as n-gram, NER, topic modeling, and emotions.

## References

- Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.
- Scott R Baker, Nicholas Bloom, Steven J Davis, and Marco C Sammon. 2021. What triggers stock market jumps?
- Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale covid-19 twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324.
- Rajeev Bhargava, Xiaoxia Lou, Gideon Ozik, Ronnie Sadka, and Travis Whitmore. 2022. Quantifying narratives and their impact on financial markets. *Available at SSRN 4166640*.
- Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Jaroslav Bukovina. 2016. Social media big data and capital markets—an overview. *Journal of Behavioral and Experimental Finance*, 11:18–26.
- Julien Chaisse and Cristen Bauer. 2018. Cybersecurity and the protection of digital assets: assessing the role of international investment law and arbitration. *Vand. J. Ent. & Tech. L.*, 21:549.
- Wallace Chipidza, Elmira Akbaripouribazar, Tendai Gwanzura, and Nicole M Gatto. 2022. Topic analysis of traditional and social media news coverage of the early covid-19 pandemic and implications for public health communication. *Disaster medicine and public health preparedness*, 16(5):1881–1888.
- Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjali Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4774–4778. IEEE.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Casey Dougal, Joseph Engelberg, Diego Garcia, and Christopher A Parsons. 2012. Journalists and the stock market. *The Review of Financial Studies*, 25(3):639–679.
- Joseph E Engelberg and Christopher A Parsons. 2011. The causal impact of media in financial markets. *the Journal of Finance*, 66(1):67–97.
- Lily Fang and Joel Peress. 2009. Media coverage and the cross-section of stock returns. *The journal of finance*, 64(5):2023–2052.
- Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 216–223.
- John Griffith, Mohammad Najand, and Jiancheng Shen. 2020. Emotions in the stock market. *Journal of Behavioral Finance*, 21(1):42–56.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Omair Haroon and Syed Aun R Rizvi. 2020. Covid-19: Media coverage and financial markets behavior—a sectoral inquiry. *Journal of Behavioral and Experimental Finance*, 27:100343.
- Gaurav Jariwala, Harshit Agarwal, and Vrai Jadhav. 2020. Sentimental analysis of news headlines for stock market. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–5. IEEE.
- Shimon Kogan, Tobias J Moskowitz, and Marina Niessner. 2021. Social media and financial news manipulation. *Available at SSRN 3237763*.
- Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.
- Gaye-Del Lo, Isaac Marcelin, Théophile Bassène, and Babacar Sène. 2022. The russo-ukrainian war and financial markets: the role of dependence on russian commodities. *Finance Research Letters*, 50:103194.
- Paul Luff and Christian Heath. 2012. Some ‘technical challenges’ of video analysis: social actions, objects, material realities and the problems of perspective. *Qualitative Research*, 12(3):255–279.
- Allan M Malz. 2021. The gamestop episode: What happened and what does it mean? *Journal of Applied Corporate Finance*, 33(4):87–97.

- Mark K McBeth, Robert J Tokle, and Susan Schaefer. 2018. Media narratives versus evidence in economic policy making: The 2008–2009 financial crisis. *Social Science Quarterly*, 99(2):791–806.
- Leela Mitra and Gautam Mitra. 2011. Applications of news analytics in finance: A review. *The handbook of news analytics in finance*, pages 1–39.
- László Nemes and Attila Kiss. 2021. Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, 5(3):375–394.
- Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. bert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7047–7055.
- Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, page 673.
- Jingwei Piao. 2015. Financial media, globalisation, and china’s economic integration: comparing narrative construction of the economist and caijing.
- Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.
- Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.
- Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- Robert P Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. 2012. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464.
- Chareen Snelson, Dazhi Yang, and Torrence Temple. 2021. Addressing the challenges of online video analysis in qualitative studies: A worked example from computational thinking research. *The Qualitative Report*, page 1974.
- Joanna Strycharz, Nadine Strauss, and Damian Trilling. 2018. The role of media coverage in explaining stock market fluctuations: Insights for strategic financial communication. *International Journal of Strategic Communication*, 12(1):67–85.
- Marc Velay and Fabrice Daniel. 2018. Using nlp on news headlines to predict index trends. *arXiv preprint arXiv:1806.09533*.
- Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433.
- Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.
- Piotr Żuk and Paweł Żuk. 2020. Right-wing populism in poland and anti-vaccine myths on youtube: Political and cultural threats to public health. *Global Public Health*, 15(6):790–804.