Textual Analysis Group Exam

The NVIDIA stock in Bloomberg News during 2016-2024

Applied Textual Data Analysis for Business and Finance

BAN432

Norges Handelshøyskole, Høst 2024

Candidate Number: 11, 23, 40, 54

Table of Contents

Task 1. Describe the corpus

For this assignment we have created three different corpora derived from the raw data. The raw data was transcripts from Bloomberg business news channel during the years 2016-2024. For task 2 we created Corpus 1. This was the raw data with transcripts from individual TV shows. We filtered for the shows mentioning Nvidia. Then we concatenated this to daily content, cleaned the documents for stopwords, numbers, punctuation, and words smaller than 2 or larger than 20 characters. We then transformed the dataset into a corpus and a document-term-matrix. By using a dataset consisting of an aggregate of whole episodes mentioning Nvidia, we hoped to also capture related companies, that were similar to Nvidia, while still keeping the corpora based on Nvidia being mentioned. For task 3 we created a second corpus, Corpus 2, by reducing Corpus 1 even further, aiming to capture the sentiment around Nvidia. We extracted strings with the term 'nvidia' or 'nvidias' and 50 words to the left and right (up to 101 words per string). Then we filtered for words, that had been used 4 or less times, to rule out misspellings or other rare words. We sum the counts for relevant versions of Nvidia (*nvidia* and *nvidias*) as a daily frequency of media attention. We also filtered out few redundant words with no meaning for the text, such as metadata. For task 4 and 5 we created a third corpus, Corpus 3, based upon Corpus 1. For Corpus 3 we filtrated strings with 'nvidia' or 'nvidias' and up to 25 words to the left and to the right, hoping to catch the most relevant and specific topics related to Nvidia. In table 1 we show the summary statistics for each of the three corpora:

Corpus 1 (Task 2)		Corpus 2 (Task 3)		Corpus 3 (Task 4 and 5)	
Metric	**Value**	**Metric**	**Value**	**Metric**	**Value**
Nr. of Documents	929	**Nr. of Documents**	897	**Nr. Of Documents**	898
Total Words	10137679	**Total Words**	109059	**Total Words**	76483
Unique Terms	30894	**Unique Terms**	11719	**Unique Terms**	9597
Avg. words per doc	10912	**Avg. words per doc.**	122	**Avg. words per doc.**	85
Top terms	**Total count**	**Top terms**	**Total count**	**Top ten terms.**	**Total count**
will	132207	**nvidia**	3710	**nvidia**	4693
us	85532	**market**	1007	**market**	678
think	76660	**tech**	742	**tech**	515
going	64302	**stocks**	653	**chips**	499
see	60450	**company**	636	**earnings**	487
get	57502	**companies**	618	**apple**	487
year	56949	**earnings**	599	**stocks**	461
market	55951	**chips**	582	**companies**	454
can	54488	**lot**	529	**scanned**	446
now	49588	**china**	525	**company**	429

Table 1. *The three corpora and their description.*

The three corpora are increasingly narrow in their filtration and the average word per document. In Corpus 1 it is on average 10912 words per document with a total of 10.137.679 words. In Corpus 2 it is 122 words on average per document with a total of 109.059 words. In Corpus 3 it was 85 words on average per document and a total of 76.483 words.

In Corpus 1 we notice that the content has several top keywords, that does not display Nvidia or expected Nvidia related words. We chose this to allow for complexity in the word embedding model using cosine similarity. In Corpus 2 and 3, Nvidia becomes the most frequent term, implying a more specific selection of data, used for the sentiment analysis in task 2, and the n-gram analysis and topic model in task 4 and 5.

Task 2. Word embedding and cosine similarity

For task 2 we identify the most similar companies to Nvidia as a competitor analysis. Using Corpus 1 for a word embedding model using cosine similarity we find the most similar companies in the corpus to Nvidia. First, we report the overall similar companies for the whole time period 2016-2024 (Table 2) and then how they change over the years, estimated annually with the data for one year at a time (Plot 1). We identified companies from the SP500 index, including the older names Facebook and Google for Meta and Alphabet. We also included OpenAI. We were not able to find a free historical index, and therefore we used the present-day index, as provided by the *tidyquant* package in R (Dancho, M., & Vaughan, D., 2024).

Based on the whole period 2016-2024, the most similar companies were:

Term	Similarity	Ticker	Company
apple	0,836	AAPL	APPLE INC
intel	0,831	INTC	INTEL CORP
tesla	0,780	TSLA	TESLA INC
microsoft	0,776	MSFT	MICROSOFT CORP
amazon	0,712	AMZN	AMAZON.COM INC
micron	0,710	MU	MICRON TECHNOLOGY INC
meta	0,696	META	META PLATFORMS INC CLASS A
alphabet	0,696	GOOGL	ALPHABET INC CL A & CL C
qualcomm	0,620	QCOM	QUALCOMM INC
broadcom	0,610	AVGO	BROADCOM INC

Table 2. *SP500 companies (Nov 2024) mentioned in the corpus ranked by cosine similarity. Estimated with Corpus 1.*

Throughout the period these companies changed, as well as the terms. Following is a barplot with the top 5 most similar companies for each year:

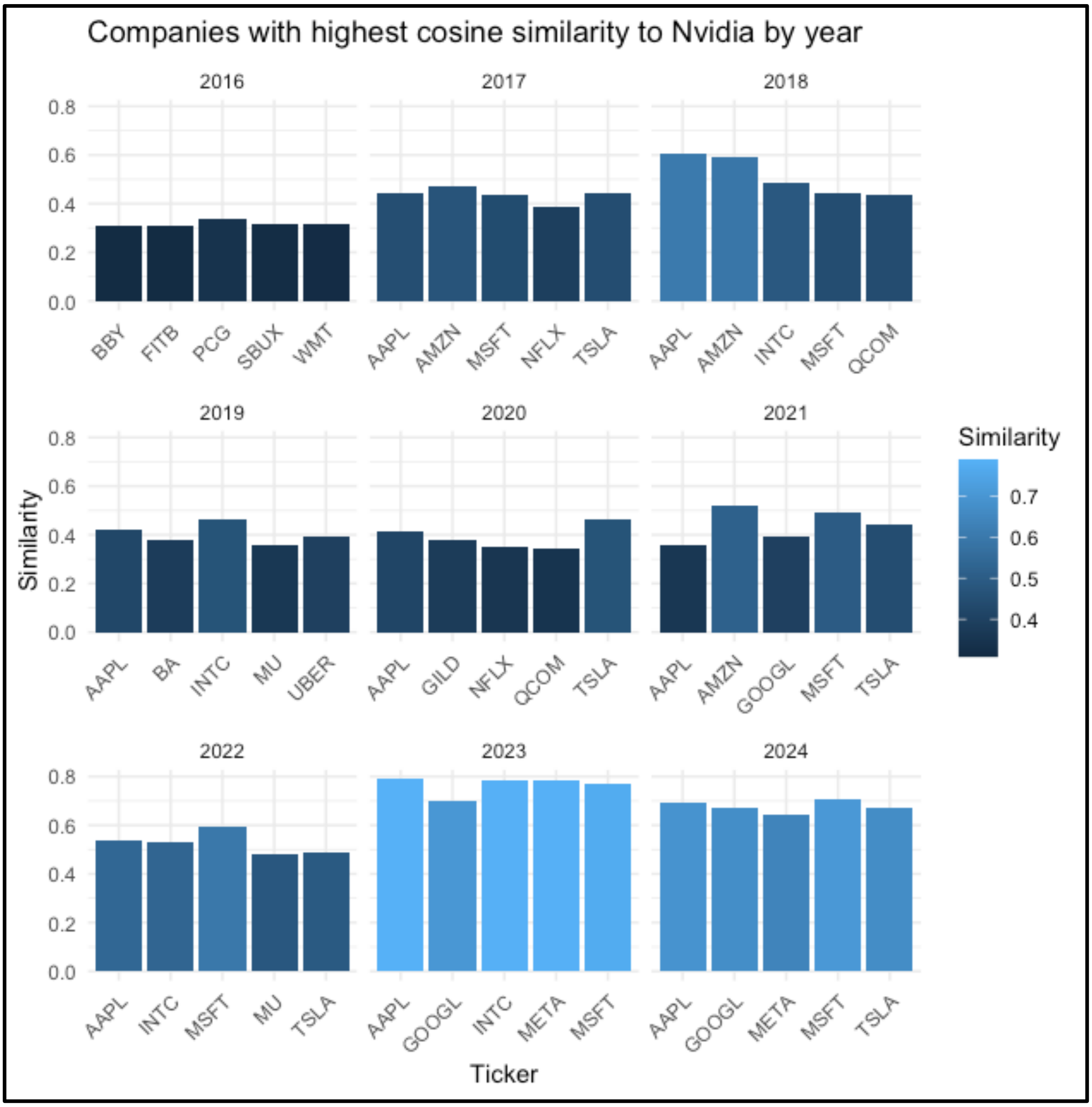


Figure 1: Top five cosine similar companies estimated annually with annual data.

In figure 1 we notice how the similarity changes for i.e. Apple (APPL) over the years, with a peak in similarity during 2023, however present for 8 years 2017-2024. Estimated with Corpus 1. Apple was amongst the top five similar companies 8 out of 9 years.

We also notice a change in the companies that are similar to Nvidia throughout the years. From companies with a relatively low cosine similarity in 2016 (like Best Buy (BBY) and Public Consulting Group (PCG), to being strongly similar to several of the largest growth companies in the SP500 index during 2023 and 2024, such as Apple (AAPL), Alphabet (GOOGL), Microsoft (MSFT), and Tesla (TSLA), and several chip manufacturers, such as Intel Corp. (INTC), Qualcomm (QCOM). Tesla (TSLA) and Meta are also similar, however more likely related to Nvidia as customers not competitors. For this assignment we have shown what other firms are related to or most similar to Nvidia and how the most similar companies changed throughout the years.

Task 3. Sentiment, quantity and stock return

For this assignment we chose to use Corpus 2 to try to better capture the sentiment associated with Nvidia. We performed a sentiment analysis using three sentiment lexica and then compared this with the daily frequency of Nvidia being mentioned in the Bloomberg News, and the daily stock price changes (i.e. it's return in %) in a correlation analysis, and a linear regression on the daily return.

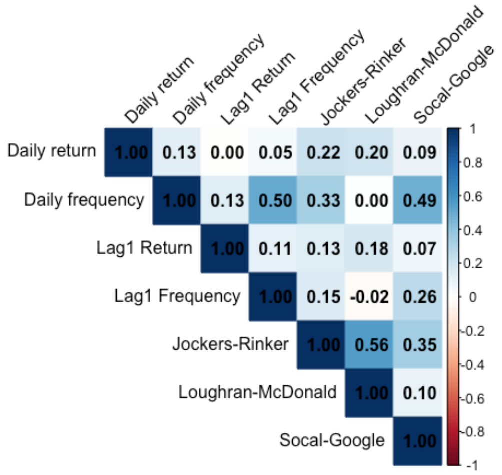
The three different lexica were:

- Jockers-Rinker (More general sentiment lexicon, ranges between –2 to 1)
- Loughlan-McDonald (Financial sentiment lexicon, ranges between –1 to 1)
- Socal-Google (Emotional intensity of words, ranges from –30 to 30).

The lexica are different in that Jockers Rinker tries to score sentiment in general text, Loughlan-McDonal scores financial reporting sentiment, and Socal-Google scores the use of emotionally charged words (Rinker et al, 2019).

Correlation

First, we estimate correlation between the daily estimates of sentiment, return and frequency, including the last days lagged return and lagged frequency:



	Daily return	Daily frequency	Lag1 Return	Lag1 Frequency	Jockers-Rinker	Loughran-McDonald	Socal-Google
Daily return	1.00	0.13	0.00	0.05	0.22	0.20	0.09
Daily frequency		1.00	0.13	0.50	0.33	0.00	0.49
Lag1 Return			1.00	0.11	0.13	0.18	0.07
Lag1 Frequency				1.00	0.15	-0.02	0.26
Jockers-Rinker					1.00	0.56	0.35
Loughran-McDonald						1.00	0.10
Socal-Google							1.00

Figure 2: Correlation matrix.

From the correlation matrix in figure 2, we notice a slight correlation between the daily frequency of mentioning Nvidia, and the sentiment scores Jockers-Rinker and Social-Google. We also notice that the daily return is slightly positively correlated with Jockers-Rinker (0,22) and Loughran-McDonald (0,20). We also notice that the daily frequency is positively correlated with past days frequency (0.5) and slightly to Dailly return (0.13).

From a quantitative point of view se can see that the daily return is slightly, but positively, correlated with daily frequency, Jockers Rinker sentiment and Loughlan-McDonald sentiment scores. Social google only has a low to neutral correlation (0.09).

Linear regression

We have tried to estimate the importance of the correlation between stock return and the frequency and sentiment scores using a linear regression. However, the relationship, such as causality and potential interdependence or biased variables, would require a deeper analysis. Return could be dependent upon sentiment and media attention, while both sentiment and attention certainly also could be dependent upon return.

We get the following linear relationship between the daily return of the Nvidia stock as a function of the daily sentiment scores and frequency of Nvidia being mentioned (i.*e. daily word count)*:

Linear regression on *Daily return* (dependent variable) using Corpus 2.
Including interaction terms between daily frequency and sentiment.

Term	Estimate	Std.error	t statistic	P value	Sign. Level
(Intercept)	0,00067	0,0030	0,23	0,82	
Daily frequency	0,00044	0,0007	0,64	0,52	
Jockers-Rinker	0,02834	0,0061	4,64	0,00	***
Loughran-Mcdonald	-0,00936	0,0070	-1,33	0,18	
Social-Google	-0,00405	0,0023	-1,75	0,08	.
Daily frequency: Jockers-Rinker	-0,00269	0,0009	-2,95	0,00	**
Daily frequency: Loughran-Mcdonald	0,00456	0,0011	4,20	0,00	***
Daily frequency: Social-Google	0,00067	0,0003	1,95	0,05	.

Table 2. Significance codes: ‘***’ = 0.001, ‘**’ = 0.01, ‘*’ = 0.05, ‘.’ = 0.1, ‘ ’ = 1.

With the risk of interdependence and bias in mind, we notice that the linear regression estimates no significant relationship between daily frequency and daily return when controlling for sentiment and interaction terms between sentiment and frequency.

We notice a significant positive relationship between Jockers-Rinker sentiment score and the return, however a negative interaction term with frequency means that this effect is reduced when frequency is increasing. This could imply that sentiment captured by Jockers-Rinker in the data is more negative on days with increased frequency, while less negative when the frequency is low.

The positive and significant interaction term between Daily frequency and Loughran McDonald indicates an enhanced effect on the return when the frequency increases. The sentiment combined with the frequency is more significant than the sentiment score by Loughran-McDonald by itself, which also makes sense, considering herd behaviour or other investor biases that could be influenced by short term price changes or financial news (Hens, 2018).

Task 4

To identify the terms associated with Nvidia’s stock performance, we analysed both unigrams and bigrams extracted from Bloomberg’s business news captions. Unigrams enabled us to detect individual significant terms, while bigrams captured commonly co-occurring word pairs that provide additional context. This approach allowed us to uncover single keywords but also meaningful phrases that correlate with periods of high and low stock prices. We used Corpus 3 (25 words before and after NVIDIA) and classified the years 2023 and 2024 as periods of high stock prices based on observations of the stock chart.

In addition to identifying key terms, we employed a log-ratio metric to quantitatively measure the association between specific unigrams and bigrams with NVIDIA's stock price categories. The log-ratio is calculated using the formula:

$$Log - Ratio = log2((FrequencyHigh + 1\ )/\ (FrequencyLow + 1))$$

where Frequency High is the count of a particular term during periods of high stock prices, and Frequency Low is its count during low stock price periods. Adding one to each frequency serves to prevent division by zero. A positive log-ratio indicates that the term is more frequently associated with high stock price periods, while a negative log-ratio suggests a stronger association with low stock price periods. This metric allowed us to differentiate and prioritize terms that were significantly more prevalent in either high or low stock price contexts.

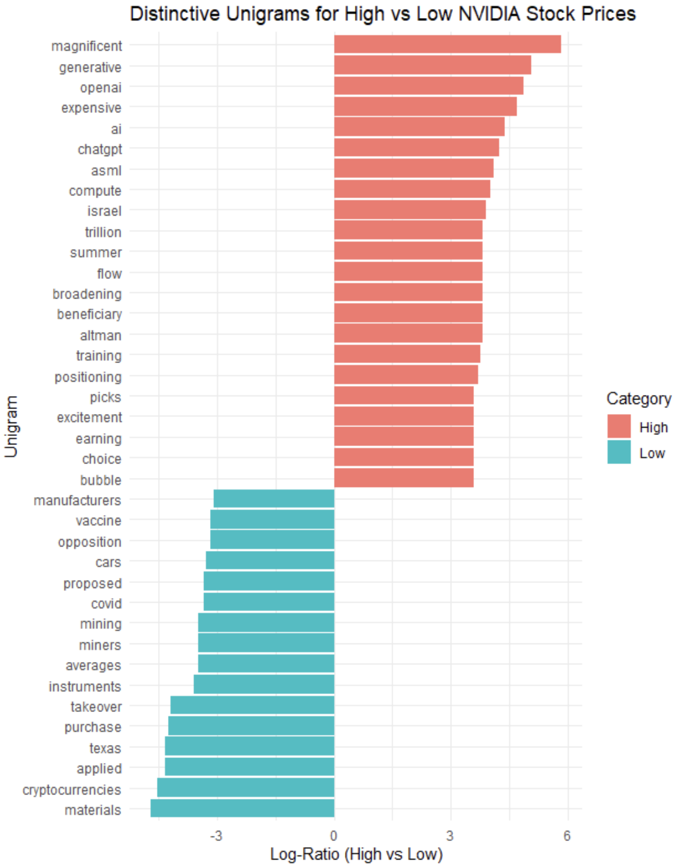


Figure 3

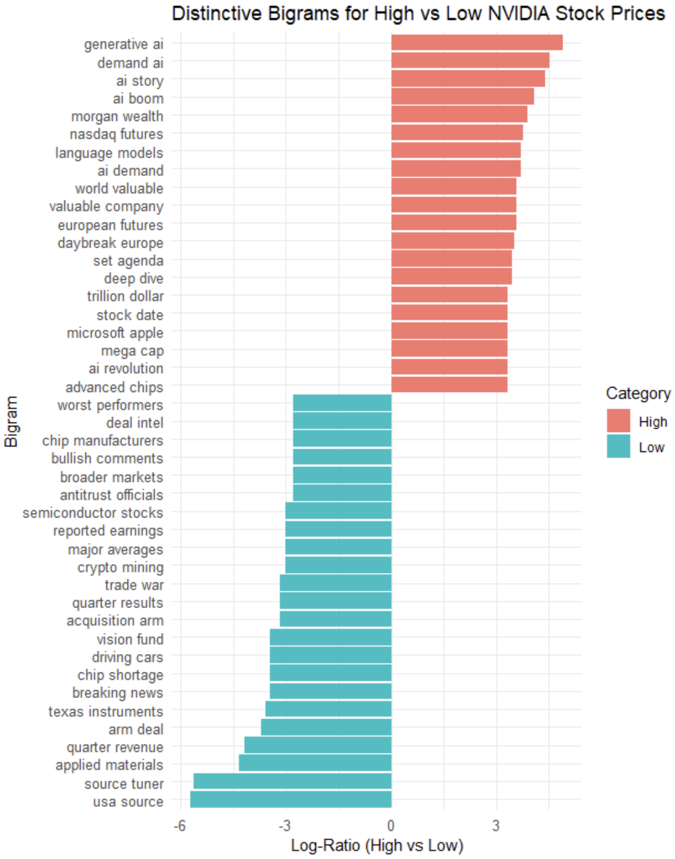


Figure 4

A possible explanation of the significant stock price increase in NVIDIA's stock price could be found in the some of the terms found in our analysis. Unigrams such as "ai", "generative", "openai", "training" and "chatgpt" is among the top distinctive unigrams during the high stock price period.

We see the same pattern in the bigrams, "demand ai", "ai story", "ai boom", "language models", "ai demand" and "ai revolution".

Nvidia is linked to the advancement of AI through developing powerful GPUs optimized for machine learning and deep learning tasks. Technologies such as the CUDA platform and Tensor Core architecture have made Nvidia's hardware indispensable for training and deploying large-scale AI models, including language models like ChatGPT. Investments in AI led to unexpected revenue growth and much higher revised revenue projections and earnings for Nvidia (Startelelogic, 2024; Nvidia, 2024).

A possible explanation for the significant increase in Nvidia's stock price can also be attributed to market hype, as indicated by our analysis of unigrams and bigrams. Terms like "bubble," "trillion," and "expensive," along with bigrams like "trillion dollar" and "stock bubble," could reflect investor enthusiasm and possible creation of fear of missing out. Conversely, our analysis identified unigrams and bigrams that highlight concerns, terms like "worst performers," "chip shortage," and "trade war" were associated with periods of lower stock performance. These terms point to external challenges like global supply chain disruptions and geopolitical conflicts impacting the semiconductor industry. These findings could be an explanation of the narrative around Nvidia's stock in "High" vs "Low" stock price. Figure 5 shows Nvidia stock price and highlights the recent period increase in price:

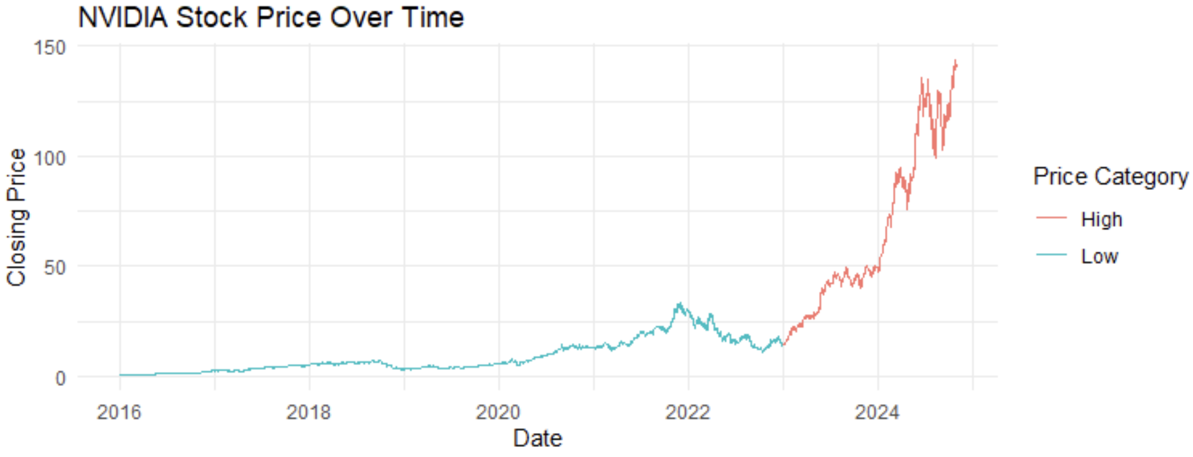


Figure 5. Nvidia stock price with markings of what we deem to be classified as "High" and "Low" prices.

Task 5

Topic modelling is a method that automatically discovers topics within a large collection of documents by identifying patterns of word usage. It groups frequently co-occurring words into clusters, revealing hidden thematic structures in the text data (IBM, 2024). By applying topic modelling to our Nvidia-related data, we uncovered the main themes discussed in the Bloomberg News, which helps us interpret factors influencing the company's stock performance.

The topic modelling implemented provides insights about Nvidia and illustrates how the topics discussed surrounding Nvidia has changed over time, shedding light on factors that may have influenced its stock price over time. Following Bar plot (figure 5) illustrates how he 8 estimated topics have changed in relative share of the discussion related to Nvidia throughout the years 2016-2024:

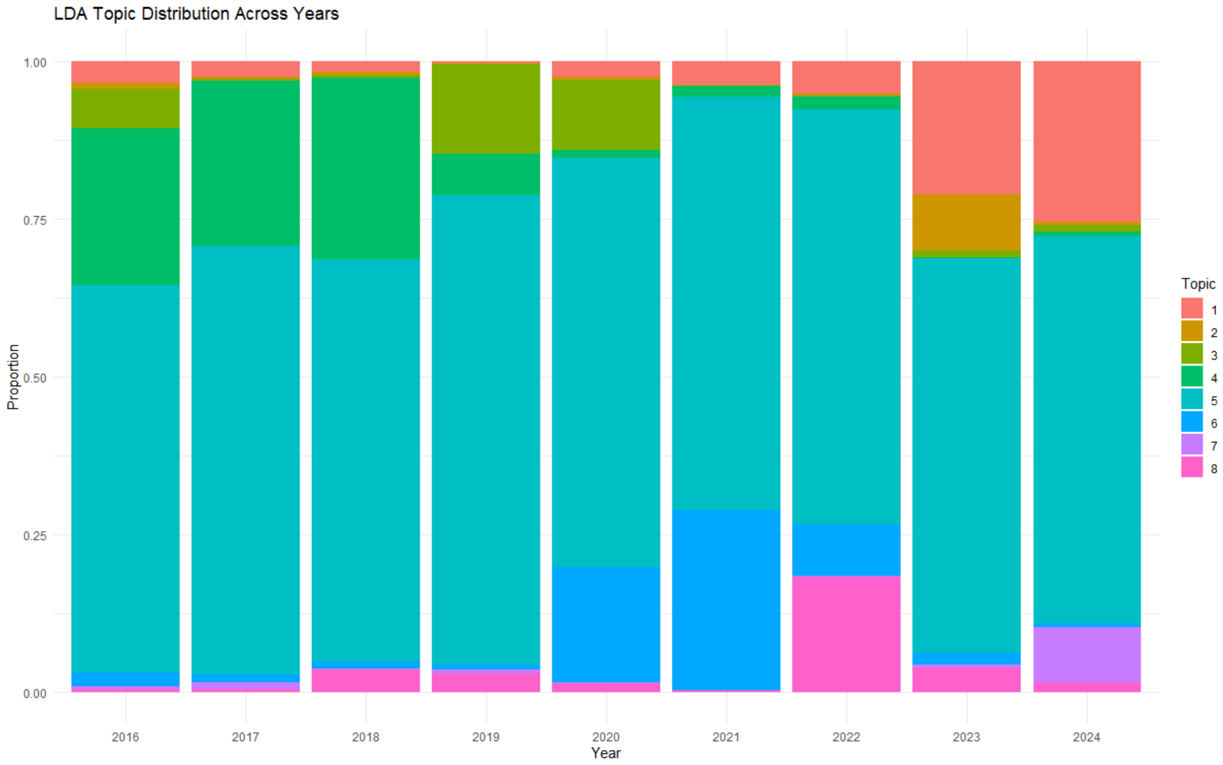
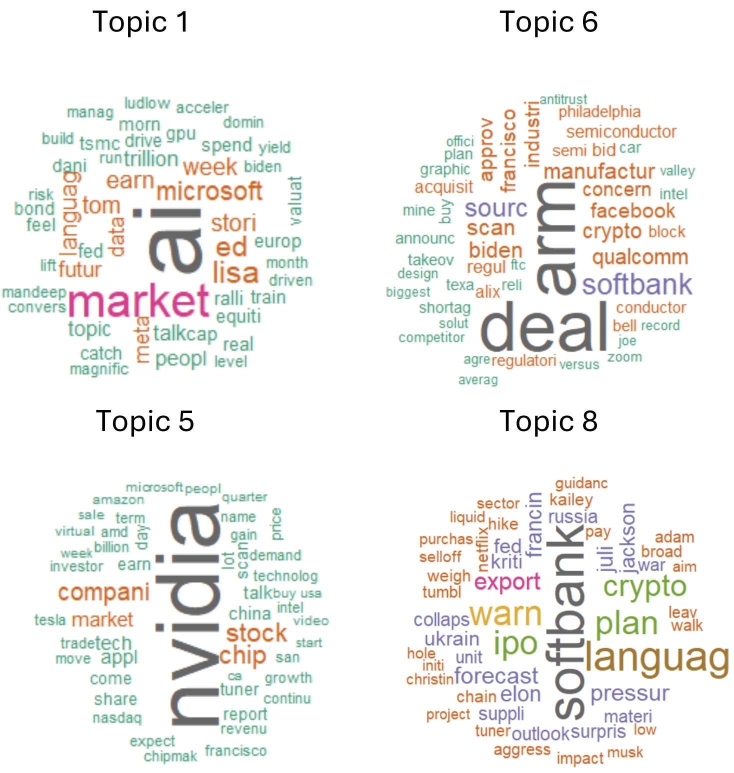


Figure 6: 8 identified topics and their relative share of the discussion on Bloomberg News related to Nvidia.

We have estimated 8 topics, and we will present the wordclouds of the five topics that were highly associated with the Nvidia during the period 2016 to 2024.



Topic 1

Topic 6

Topic 4

Topic 5

Topic 8

Figure 7. Five wordclouds of the selected topics.

Our first observation is that Topic 5 is the dominant theme in all years. Topic 5 primarily represents discussions around Nvidia, the semiconductor industry, its innovations in chip technology, stock performance, but also includes competitors and customers.

Between 2016 and 2019, Topic 4, focusing on Bitcoin, cryptocurrency, and gaming, was particularly prominent. Nvidia's GPUs were widely used for gaming and cryptocurrency mining during this period. However, cryptocurrency mining has since become a smaller part of Nvidia's business, driven by external factors and the introduction of specialized Cryptocurrency Mining Processors (CMP). This shift is reflected in the declining focus of cryptocurrency-related discussions over time. Similarly, while gaming remains a core revenue driver, it is discussed less frequently, likely due to the growing focus on other areas of Nvidia's business (NVIDIA, 2021).

Between 2020 and 2022, Topics 6 and 8 which focus heavily on Nvidia's $40 billion bid to acquire ARM from SoftBank. The deal, aimed at expanding Nvidia's reach into mobile computing and IoT, faced significant regulatory hurdles globally. (Nvidia, 2022) This is reflected in the topic models, where discussions about ARM and SoftBank were particularly prominent during this period. However, following the deal's abandonment in early 2022 and SoftBank's plans to take ARM public, these topics ceased to be discussed in relation to Nvidia.

In 2023 and 2024, the topic model reveals a sharp increase in discussions linking AI with NVIDIA, reaffirming the conclusion from Task 4 that Nvidia's part in AI technologies has been a major driver of its stock performance the past two years. NVIDIA's GPUs are essential for training and deploying advanced AI models, making the company a cornerstone of the AI revolution. This surge in AI-related demand has translated into unexpected revenue growth and higher earnings projections.

Sources:

Dancho, M., & Vaughan, D. (2024). Tidyquant: Tidy Quantitative Financial Analysis. Link: https://cran.r-project.org/web/packages/tidyquant/index.html

Hens, T., Bachmann, K., & Giorgi, E. D. (2018). *Behavioral Finance for Private Banking*. Wiley Finance, 2nd ed.

IBM. (2024). Topic Modeling. Link: https://www.ibm.com/topics/topic-modeling

Nvidia. (2021). Cryptocurrency Mining Processor (CMP) for Professional Mining. *NVIDIA Blog*. Link: https://blogs.nvidia.com/blog/geforce-cmp/

Nvidia. (2022). NVIDIA and SoftBank Group Announce Termination of NVIDIA's Acquisition of ARM Limited. *NVIDIA Newsroom*. Retrieved from https://nvidianews.nvidia.com/news/nvidia-and-softbank-group-announce-termination-of-nvidias-acquisition-of-arm-limited

Nvidia. (2024). NVIDIA Announces Financial Results for Second Quarter Fiscal 2025. Link: https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2025

Rinker et al. (2019). Lexicon: Lexicons for Text Analysis. Link: https://cran.r-project.org/web/packages/lexicon/index.html

Startelelogic. (2024). NVIDIA's Role in AI Development Powering the Future of Machine Learning. Link: https://startelelogicofficial.medium.com/nvidias-role-in-ai-development-powering-the-future-of-machine-learning-da3cb7fd0740