Topic model applications to enhance company performance

# Introduction

Following the management team’s interest in our model to predict stock market price using NLP techniques on news articles, they have shown an interest in the training of a model that is able to improve company performance using text data. To achieve this our team analyzed the contents of Nvidia related news articles and identify the topics that are present in these news articles.

To achieve this topic modelling was used for this analysis. By applying topic modelling techniques, we intend to uncover the content of the Nvidia related articles. This report relays the insights gained, inner workings and results of using three distinct topic modelling algorithms, namely Latent Dirichlet Allocation (LDA), Formal Latent Semantic Analysis Weighted (FLSA-W), and BERTopic.

In the following sections the technicalities of these algorithms will be explained, and an overview of their performance will be given. Additionally, the topic modelling outcomes under various environmental changes will also be presented. Finally, the limitations of these algorithms and the manner in which our findings could be applied to real-world financial companies will be discussed.

# The Topic Modelling Algorithms

This section explains how each of the topic models function, extract useful information from the Nvidia news articles along with their weaknesses and strengths. In knowing how these algorithms operate we can better understand their performance and how to improve them.

Latent Dirichlet Allocation (LDA)

According to a classic representation theorem by (De Finetti, 2017) any collection of exchangeable random variables has a representation as a mixture distribution—in general an infinite mixture. It is upon this theory that Latent Dirichlet Allocation (LDA) is based (David M. Blei et al., 2003). LDA is a generative statistical model that is capable of representing words and documents in a way that acknowledges their exchangeability, meaning the probability of encountering a set of words is the same regardless of their order. In other words it assumes that documents are a mixture of topics and topics are a mixture of words. Due to this capability, LDA can be utilized to determine the topics that are present in a large collection of documents. It achieves this by assigning each word in each document to a random topic and iteratively refines these assignments based on the probabilities of word occurrences in topics and topics in documents.

The strengths of LDA lie in how easy its results are to understand and interpret. LDA’s efficiency with large datasets along with its flexibility, through the incorporation of hyperparameters make it an excellent choice for the task at hand which is topic modelling. On the other hand, LDA is context ignorant, meaning it assumes that words are independent of one another(David M. Blei et al., 2003). Furthermore, LDA’s performance heavily relies on the tuning of its hyperparameters. And lastly, it may struggle with short texts. Despite these shortcomings, it strengths make LDA a solid contender for our topic models.

Fuzzy Latent Semantic Analysis Weighted (FLSA-W)

FLSA uses singular value decomposition to decompose document representations into three matrices UΣVT. It then applies importance weights (like TF-IDF) to this singular value decomposition of the term-document matrix(Rijcken et al., n.d.). This application of importance allows the model to prioritize more significant terms that occur less frequently across our corpus, but are more informative for our context, which in our case would be the Nvidia related terms. Due to this mechanism FLSA-W can effectively highlight differentiated topics and semantic relationships within Nvidia-related content by weighing terms according to their importance.

One of the traits that makes FLSA-W a prime candidate for topic modelling is that its one of the fastest and efficient topic models, due to its computational efficient manner of applying matrix operations. It is also capable of reducing noise on the most significant singular vectors and is excellent at handling synonyms. Unfortunately FLSA-W’s Dimensionality reduction might lead to the loss of important information and it’s more limited in comparison to other models when handling larger datasets.

## BERTopic

BERTopic leverages state-of-the-art transformer models to create dense vector representations of texts. It uses these embeddings, created from the BERT (embedder) to find clusters of documents, which represent the topics(Grootendorst, 2022). Unlike LDA, BERTopic considers the context and semantics of words in texts due to the transformer architecture, leading to more contextually relevant topics. It is context aware, can adjust the number of topics dynamically and performs well even on short documents(Egger & Yu, 2022). However, this prestige comes at the price of it being the most computational intense algorithms we’ve tested. It also sacrifices interpretability for performance, due to its black-box nature. Lastly this model relies heavily on the quality of the pre-trained BERT embeddings. As a result, these embeddings need to be properly trained in order for the topic model to perform optimally.

# Analysis and Approach

In this section, we will discuss the training of the selected models. Our approach consists iteratively training our topic models and slightly adapting the environment in which they are trained for every iteration. Not only does this improve our models over time but also gives us insights into how different modifications to our input data affects the model’s performance. The decisions that were made, and the setup and changes that were applied to each iteration will also be presented in this section. We chose to apply 3 iterations (excluding an iteration 0 where only some basic processing had been applied), since this allowed us to easily compare 3 versions of 3 models to one another. To further improve the comparing of these algorithms the same iteration steps were applied to each model’s iteration. More information on the iterations and their environment below.

## Iteration 0

This iteration was named ‘iteration 0’, due to us applying some basic preprocessing to the data. After filtering the out all of the Nvidia related articles, the hyperlinks were removed along with the numbers and non-alpha words. Lastly, all of the document text were represented in lowercase. Subsequently, after applying these preprocessing steps we trained the models with their default values.

## Iteration 1

For the First iterations we removed stop words (e.g.  a, the, and , or , of ) from the Nvidia related articles. Using the results gained from iteration 0 we started adjusting the number of topics and words to incorporate. By means of using elbow plots, we were able to identify the optimal number of topics based on coherence, diversity and interpretability scores (more on this later).

## Iteration 2

In the second iteration the titles of the documents were added to their respective document text. We believed this would provide some more context for the topic models. Apart from this adjustment, for the preprocessing step, lemmatization was performed, reducing the words found in the documents down to their base form (e.g. runs 🡪 run).

## Iteration 3

Lastly for the third iteration, the parameters for the models (where possible) were hypertuned. This was achieved by applying a grid search on their parameters. By applying this method of training, we can evaluate the models at their strongest and most optimal states. This will give us insights into the total capacity of these topic models in the context of extracting performance related information from the topics they generate.

# Evaluation and Results

In this section we will assess the performance of the models after each iteration. Each model was assessed qualitatively and quantitatively. For the quantitative assessment we chose to 3 metrics to compare the models. We use the coherence, diversity and interpretability score calculated by utilizing FLSA-W’s (python package called flsaw) built-in functions. The *Coherence Score* measures how well each topic’s words support each other(Rijcken et al., n.d.). A measure for inter-topic quality is topic *Diversity*, which measures the unique words in a topic model as a proportion to the total number of words (A. B. Dieng et al., 2020). The *Interpretability Score* combines intra- and inter-topic quality by taking the product between the coherence and diversity score (see the equation below)(A. B. Dieng et al., 2020). These metrics were chosen since they allow for us to objectively compare the intra-, inter-topic- and overall quality for each of our models.

*Eq. Interpretability = Coherence × Diversity*

## Qualitative Assessment

### LDA

The topics generated in iteration 0 (see table 1) appeared to be rather generic and not well-defined. There is a high overlap of common words across the generated topics. Each topic consists mostly of high frequency stop words such as "the," "to," "and," "of," "in," "on," "is," "for," and "that. This is to be expected since there were some preprocessing steps that could still be applied to produce more meaningful topics, e.g. the removing of stop words. Furthermore topics 1, 2 and 4 share a lot of similarities providing more evidence that there is a lack of distinctiveness between these topics. Another reason for this outcome could be the number of topics specified to the model (k=5) which may have been too few causing the model to group words into categories that are too broad. Lastly, since these topics consisted of mostly stop words, there were no typos to be found withing the defining words. Despite these shortcomings there is still the slight presence of Nvidia related terms and references like “Nasdaq” and “nyse”. Unfortunately, any insights that could have been gained have been diluted by the common words.

For the first iteration of training the LDA model the stop words were removed. For this iteration we can already see evidence of improvement, as the LDA model is now able to provide much more defined and relevant topics compared to the previous iteration (see Table 2). We can see for example that Topic 1 focuses on stock market related terms and daily trading activities (e.g. ‘nasdaq’, ‘stock’ and ‘trade’), while Topic 2 contains words that are more focused on describing financial metrics (e.g. ‘growth’, ‘earnings’ and ‘rank’). Topic 3 seems to lean more towards financial analysis. Topic 4 is the least specific to Nvidia related terms since these words can be found in more general market and business discussions. Finally, Topic 5 seems to focus on time frames and possibly financial performance (“year”, “quarter” and “revenues”). From the defining words, it also evident that this iteration did not produce any incorrectly spelled words. Therefore, we can conclude that this iteration has improved the quality of the topics and consist of more relevant defining words.

For the second iteration adding titles and applying lemmatization seems to have further refined the topics. They seem to be a lot more distinct and clearer in comparison to the previous iteration (see table 3). The defining words are a lot more similar (see topic 5: ‘driving’, ‘car’, ‘vehicle’). From these generated topics we can conclude that indeed this iteration of the LDA model has improved significantly due to the adjustments. These adjustments made in preprocessing have effectively enhanced the model’s ability to extract meaningful and relevant topics that may be better suited for in-depth analysis and application in various financial and technological contexts.

**The last iteration only has 2 topics???**

### FLSA-W

This section qualitative assessment is provided for the topic generated by the FLSA-W model. The most intriguing aspect of the topics generated for this model’s 0th iteration, would be the variety between topics (see Table 4). For example, Topic 1 and Topic 3 provide relative clear insights and contain some meaningful defining words, given that at this stage, only some minor preprocessing has been performed. The seem to contain words relating to Nvidia stock and general marketing respectively. On the contrary, the defining words for Topic 2 are too broad and general to provide any context related to the topic (this is the topic that contains the stop words like "the," "to," "in," "of"). The weights of these defining words also indicate that these words (specifically the stop words) do not contribute to the given topic. These weights reflect the quality of the topic. If the highest weighted words are very low, we can conclude that the defining words are not that defining, which is the case for most of these topics. Hence, we can conclude that these topics are well defined and only contain some stop words, heavily present in topic 2. This is to be expected when considering that only preprocessing has been performed, especially when comparing the quality of these topics to that of the LDA model.

When assessing the 1st iteration there is very much to note other than the removal of stop words improved the relative topics generated. Furthermore, the definition of the topics have been greatly improved by the removal of stop words (see table 5). Topic 1 for example seems to contain documents relating to market activities (trade, exchange, futures), but also contains some words that don’t really seem to belong like (e.g. sporting, bell and rose). Topic 2 has words that are names, potentially of authors or journalists (e.g. troy, thomson, ryan maybe mcqueeny) indicating that this topic might be related to financial journals and market analyses. Topics 3 seems to contain words relating to market regulation (supervisory, reiterates, delisting) Topic 4 contains the keywords more related to Nvidia and its articles as these topics contain the more relevant terms that we are familiar with, such as , nvidia, nvda, technologies. The difference between topics however is still a small issue, seeing that keywords such as ‘nasdaq’ appear in multiple topics.

The topics generated in the third iteration of the FLSA-W model, with lemmatization and the inclusion of document titles, appear to reflect more specific topics (see table 6). For the first topic financial market movements seem to be the theme based on the defining words (various abbreviations for entities like ‘phlx’, ‘amat’ and ‘bac’). Overall, the application of lemmatization and the addition of document titles have refined the granularity of the topics, leading to clearer and more precise topics. We can conclude that this iteration was particularly effective at isolating distinct themes relevant to business and technology.

For the final iteration of the FLSA-W topic model, where the model has been hypertuned to obtain the optimal parameters, the models generated demonstrate some more coherence than past iterations. The diversity between the topics, despite having varying keywords, seem to correctly depict that all of these topics are associated with financial analysis (topics 1 and 5), market dynamics and technology adoption (e.g. topic 2). The goal of this analysis was to identify topics related to ‘financial’ performance. If we focus on topic 3, it is clear that out of all the iterations this topic has the closest approximation and greatest potential to achieve this end. With words like 'surged', 'gained' indicating stock movement and ('investing', 'boost') denoting potential investment strategies, this topic is deemed the most insightful with regards to financial performance.

Quantitative Assessment

# Limitations

As we can see from the results (see Appendix) the three topic models in question are capable of generating meaningful topics that are relatively coherent and well defined. However, the goal of this analysis was to improve company performance using text data. Even the most refined iterations and best version of our models, despite their valuable insights have limitation and are constrained not only by the task for which they were being created, but also limitations that apply to topic modeling in general. In this section these limitations will be discussed.

***Scalability and Computational Cost:***

Topic models can become computationally expensive and less effective as the size and complexity of the dataset increase, this limits our approach in the sense that the dataset needs to be balanced(Alsumait et al., 2008). It needs to be large enough so that the models can properly learn, but not too large that the models can actually produce output. Models like BERTopic would take too long to run if the dataset would be too large.

***Data Quality and Preparation****:*

The performance of topic models depends very highly on the quality of the text data and how well the preprocessing steps are conducted (Schofield et al., 2017). This was especially evident when attempting to train the first iteration of the BERTopic model without removing stop words. The sheer volume of extra words made it impossible for the model to generate topics within reasonable time (model ran for about 20 minutes without showing any signs of completing).

***Context and Nuance****:*

Topic models are limited in their ability to capture the nuances and context of language. Improvements in iterations may help manage context to some extent, for example, through better preprocessing or more advanced modeling (like integrating BERTopic which utilizes embeddings). However, capturing the full scope of linguistic nuances, especially in complex business or technical discussions, goes beyond what statistical topic models can achieve(A. Dieng et al., 2020).

These are the limitations that hinder topic models. Not just in their ability to be used in other applications like predicting the stock market movement or improving a company's performance, but in general topic modelling itself. Unfortunately, improving the models through various iterations and only taking the final and best iterations of each model into consideration, these limitations still apply. Topic modeling techniques inherently carry certain constraints regardless of the iterations or refinements made during the modeling process especially when their output is intended for more strategic applications.

# Conclusion (how can your topic model be applied in a financial company)

# References

Alsumait, L., Barbara, D., & Domeniconi, C. (2008). *On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking* (p. 12). https://doi.org/10.1109/ICDM.2008.140

David M. Blei, Andrew Y. Ng, & Andrew Y. Ng. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

De Finetti, B. (2017). *Theory of probability: A critical introductory treatment* (A. Machì & A. F. M. Smith, Trans.). John Wiley & Sons.

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453. https://doi.org/10.1162/tacl\_a\_00325

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (No. arXiv:2203.05794). arXiv. http://arxiv.org/abs/2203.05794

Rijcken, E., Zervanou, K., Mosteiro, P., Spruit, M., Scheepers, F., & Kaymak, U. (n.d.). *A Performance Evaluation of Topic Models based on Fuzzy Latent Semantic Analysis*.

Schofield, A., Thompson, L., & Mimno, D. (2017). *Quantifying the Effects of Text Duplication on Semantic Models*. https://doi.org/10.18653/v1/D17-1290

# Appendix

|  |  |
| --- | --- |
| **Topic Number** | **Defining words** |
| Topic 1 | the to nasdaq at nyse on was inc and in |
| Topic 2 | the to and of in that is on for it |
| Topic 3 | the of and in to for on is as that |
| Topic 4 | the of to and in zacks is year company for |
| Topic 5 | the to and of in that is for with its |

*Table 1. This table depicts the topics and the words that defined each topic respectively for iteration 0 of the LDA model. It shows that there is a lack of distinctiveness.*

|  |  |
| --- | --- |
| **Topic Number** | **Defining words** |
| Topic 1 | nasdaq stock nyse inc nvidia shares stocks day nvda trade |
| Topic 2 | nasdaq nyse trade china year index stocks shares earnings percent |
| Topic 3 | zacks earnings year stocks company growth rank stock quarter nasdaq |
| Topic 4 | market said one could companies week also like last new |
| Topic 5 | year company nvidia quarter nasdaq million billion zacks revenues intel |

*Table 2. This table depicts the topics and the words that defined each topic respectively for iteration 1 of the LDA model. It shows that there is an improvement in the topic quality with regards to relevance compared to iteration 0.*

|  |  |
| --- | --- |
| **Topic Number** | **Defining words** |
| Topic 1 | point percent close trading fell rose dow oil future friday |
| Topic 2 | current portfolio past performance today security value firm return eps |
| Topic 3 | fiscal reported cent positive solution cloud demand margin corporation gaap |
| Topic 4 | etf fund traded past read holding volume exposure since global |
| Topic 5 | driving car google self ai vehicle amazon facebook autonomous device |

*Table 3. This table depicts the topics and the words that defined each topic respectively for iteration 2 of the LDA model. It shows that there is an even greater improvement in the topic quality, and it is also evident that the within-topics defining words are much closely related to one another.*

|  |  |  |
| --- | --- | --- |
| **Topic Number** | **Defining words** | **Weights** |
| Topic 1 | nvda | 0.0019 |
| nvidia | 0.0019 |
| com | 0017 |
| nasdaq | 0.0016 |
| shares | 0.0015 |
| nyse | 0.0015 |
| advanced | 0.0015 |
| days | 0.0015 |
| during | 0.0015 |
| p | 0.0015 |
| Topic 2 | the | 0.0002 |
| to | 0.0002 |
| in | 0.0002 |
| of | 0.0002 |
| a | 0.0002 |
| and | 0.0002 |
| on | 0.0002 |
| for | 0.0002 |
| s | 0.0002 |
| with | 0.0002 |
| Topic 3 | traded | 0.0015 |
| reuters | 0.0015 |
| exchange | 0.0015 |
| composite | 0.0015 |
| session | 0.0015 |
| chipmaker | 0.0008 |
| rose | 0.0007 |
| blast | 0.0007 |
| previously | 0.0006 |
| bell | 0.0006 |
| Topic 4 | book | 0.009 |
| evolution | 0.0077 |
| books | 0.0075 |
| richard | 0.0071 |
| science | 0.0066 |
| certainty | 0.006 |
| modern | 0.0056 |
| professor | 0.0055 |
| me | 0.005 |
| beliefs | 0.0045 |
| Topic 5 | gmt | 0.0005 |
| outnumbered | 0.0004 |
| performers | 0.0003 |
| settle | 0.0003 |
| the | 0.0003 |
| to | 0.0003 |
| slumped | 0.0003 |
| in | 0.0003 |
| of | 0.0003 |
| a | 0.0003 |

*Table 4. This table illustrates the topics generated for FLSA-W 0th iteration. There is some topics that provide useful insights while others consist of some if not entirely of stop words.*

|  |  |  |
| --- | --- | --- |
| **Topic Number** | **Defining words** | **Weights** |
| Topic 1 | reuters | 0.001 |
| traded | 0.001 |
| exchange | 0.0008 |
| composite | 0.0008 |
| session | 0.0008 |
| bell | 0.0008 |
| futures | 0.0008 |
| rose | 0.0008 |
| sporting | 0.0008 |
| chipmaker | 0.0008 |
| Topic 2 | towards | 0.0006 |
| bidu | 0.0005 |
| pharmaceuticals | 0.0005 |
| cures | 0.0005 |
| philadelphia | 0.0005 |
| uncover | 0.0005 |
| miners | 0.0005 |
| semis | 0.0004 |
| ounce | 0.0004 |
| unchanged | 0.0004 |
| Topic 3 | gmt | 0.0005 |
| troy | 0.0004 |
| settle | 0.0004 |
| premarket | 0.0003 |
| implied | 0.0003 |
| thomson | 0.0003 |
| nellis | 0.0003 |
| bartosiak | 0.0003 |
| ryan | 0.0003 |
| mcqueeney | 0.0003 |
| Topic 4 | nasdaq | 0.0002 |
| also | 0.0002 |
| market | 0.0002 |
| reiterates | 0.0002 |
| new | 0.0002 |
| delisting | 0.0002 |
| company | 0.0002 |
| alv | 0.0002 |
| supervisory | 0.0002 |
| equitiesasian | 0.0002 |
| Topic 5 | com | 0.0002 |
| nvda | 0.0019 |
| nvidia | 0.0018 |
| believe | 0.0018 |
| advanced | 0.0017 |
| intc | 0.0017 |
| broader | 0.0017 |
| days | 0.0016 |
| nasdaq | 0.0016 |
| technologies | 0.0016 |

*Table 5. This table illustrates the topics generated for FLSA-W 1st iteration. We can conclude that removing stop words has improved the overall quality and potential for providing meaningful content.*

|  |  |  |
| --- | --- | --- |
| **Topic Number** | **Defining words** | **Weights** |
| Topic 1 | inbox | 0.0006 |
| previously | 0.0005 |
| ratio | 0.0005 |
| amat | 0.0005 |
| ltd | 0.0005 |
| contract | 0.0005 |
| ltd | 0.0005 |
| advancing | 0.0005 |
| bac | 0.0005 |
| gmt | 0.0005 |
| Topic 2 | generation | 0.0182 |
| margin | 0.0161 |
| expect | 0.0156 |
| executive | 0.0144 |
| gross | 0.0143 |
| know | 0.0142 |
| half | 0.0136 |
| think | 0.0135 |
| chief | 0.013 |
| server | 0.013 |
| Topic 3 | premarket | 0.0004 |
| unveils | 0.0003 |
| smh | 0.0003 |
| uncover | 0.0003 |
| conclusively | 0.0003 |
| esp | 0.0003 |
| avoided | 0.0003 |
| soxx | 0.0003 |
| strategiesit | 0.0003 |
| delve | 0.0003 |
| Topic 4 | compared | 0.0033 |
| adoption | 0.0033 |
| gaming | 0.0032 |
| basis | 0.0031 |
| computing | 0.0031 |
| ago | 0.0029 |
| segment | 0.0028 |
| believe | 0.0028 |
| significant | 0.0027 |
| fourth | 0.0027 |
| Topic 5 | corporation | 0.0017 |
| investing | 0.0016 |
| corp | 0.0015 |
| gained | 0.0015 |
| com | 0.0015 |
| surged | 0.0014 |
| released | 0.0014 |
| point | 0.0014 |
| added | 0.0013 |
| boost | 0.0013 |

*Table 6. This table illustrates the topics generated for FLSA-W 2nd iteration. We can conclude that the addition of titles and applying lemmatization has enabled the model to provide more specific topics.*