FACULTY OF SCIENCE, ENGINEERING AND ENVIRONMENT

ASSESSMENT REPORT (NATURAL LANGUAGE PROCESSING)

Topic Modelling of Financial Discussions on Reddit using Latent Dirichlet Allocation LDA and BERTopic

SUBMITTED BY

JOSHUA EMMANUEL EVUETAPHA (@00790872)

DATE

23-04-2024

INTRODUCTION

Topic Modelling is a Natural Language Processing(NLP) technique used to organise, summarise and understand a large amount of text documents by grouping them into topics. Topic modelling allows us to get the central theme across documents without needing a label. It is an unsupervised learning technique, and the number of topics needs to be decided by the user.

For this task, I am using posts from Reddit. Reddit is a popular social media platform and online forum where users share, discuss, and vote on content across different niche communities known as subreddits. These subreddits have rules on the type of content permitted within them. There are thousands of subreddits, which makes Reddit a good source for textual data and rich discussions across various subjects. For this task, I used recent posts about finance and investments obtained from popular subreddits about finance, stocks, crypto, and tax. We can identify common topics by using topic modelling on these large volumes of text.

The Topic modelling of posts on top financial subreddits will help firms and individuals get insights into the recent discussion around finance and investments, which can help make sound financial and investment decisions or market financial instruments.

Latent Dirichlet Allocation (LDA): LDA is a probabilistic model. It assumes that documents (posts) are made up of a mixture of topics, and those topics are generated from words based on their occurrence. (Bansal, 2025). For example, words like "stocks", "market", "investments", "money" and "trading" might appear a lot in posts related to finance and investment instruments.

BERTopic: BERTopic is based on the popular language model BERT (Bidirectional Encoder Representations from Transformers), which is trained on a large corpus of text data to understand the context and meaning of words. Unlike LDA, which is based solely on word probabilities and co-occurrence, BERTopic uses language embeddings to understand the context and relationships between words. For example, in finance-related posts, words like "stocks", "market", "investments", "money", and "trading" might appear in different discussions.

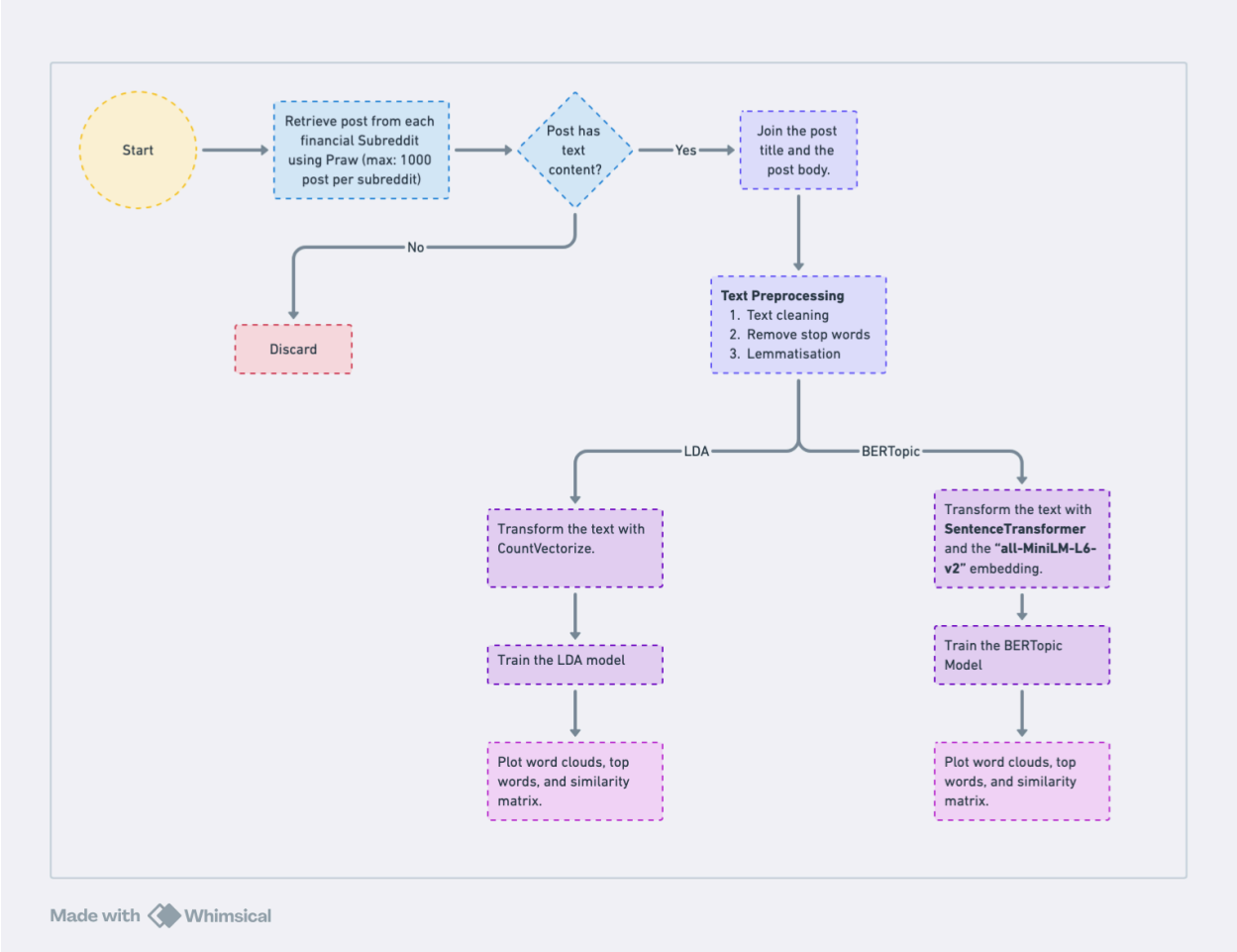


Figure 1

DATA COLLECTION

The datasets for this task were obtained from over twenty subreddits about finance and investments with the help of the reddit api and the **praw** Python package.

The data collection script queried the latest one thousand posts from each subreddit, but usually returns less than that due to some posts having no text content. The one thousand post limit is also a hard limit set by the Reddit API. This limit stopped me from getting more data per subreddit or looking far back for old posts.

The post data was obtained from the following subreddits: finance, investing, stocks, wallstreetbets, StockMarket, economy, ValueInvesting, personalfinance, fire, FinancialPlanning, UKPersonalFinance, AusFinance, EuropeFIRE, CryptoCurrency, Bitcoin, RealEstateInvesting, Daytrading, Options, Forex, AlgorithmicTrading, Tax, Dividends.

TEXT PREPROCESSING

The following steps were taken to clean and prepare the text data for the topic modelling algorithm:

1. **Joining the title and the text column**: The data from Reddit has two text columns, title and text. I joined these two columns together to create a third column called content. This was done because the title contains vital textual information valuable to a topic modelling algorithm.
2. **Text cleaning:** The data contains irrelevant information, like links, special characters, punctuation, emojis, etc. These pieces of information could affect our analysis poorly, so they need to be removed. After cleaning the data, we convert every text in our corpus to lowercase. This step is essential to prepare our data for the Topic Modelling algorithm.
3. **Stopwords removal:** Stopwords are common English words used in most texts, but they have little or no meaning.
4. **Lemmatisation:** This is reducing the word to its base dictionary form called lemma, e.g run, ran and running are different words, but they have the same base word "run". With lemmatisation, they will all be reduced to run. This will improve our models, especially LDA. I used WordNet from NLTK for lemmatisation. Both models will use the lemmatised text for the analysis.
5. **Data Preparation for each model (Text Representation)**: LDA and BERTopic expect their data to come in different formats.

EXPLORATORY ANALYSIS

Total Number of Unique Tokens

The total number of tokens gives us an Idea of our total corpus size, which will aid in selecting the max feature during training. The number of unique tokens for the raw text after cleaning was 48,181; after stopword removal, it reduced to 48,050 and finally to 44595.

Text Distribution

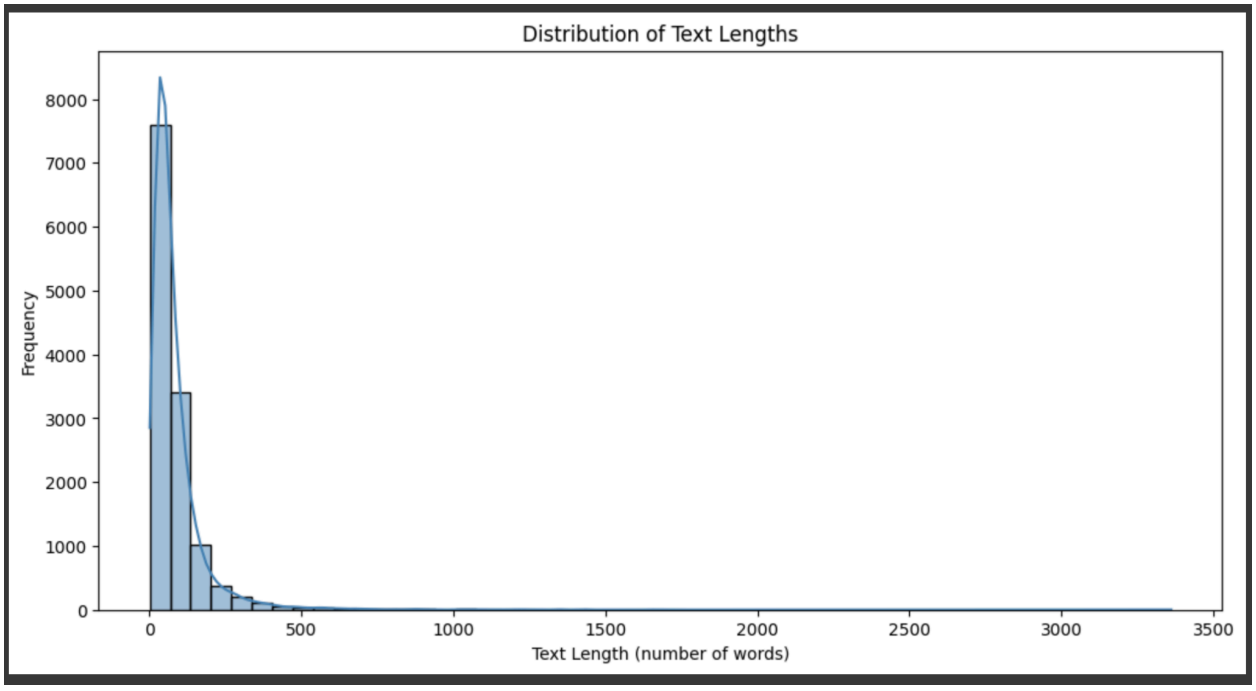


Figure 2

The histogram above is right-skewed; most posts are between 1 and 150 words long. This is consistent with social media posts since the data was obtained from Reddit. However, it has a long tail with a few posts over 1000 words, which are outliers and posts with high-quality discussion.

Further analysis showed that about 341 posts had more than 300 words, about 119 posts had more than 500 words, 24 posts had over 1000 words, four posts had over 2000 words, and only one post had over 3000 words. After conducting more analysis on these outliers by reading them, I discovered that they had a financial theme and because of that, I won't be dropping them.

Wordcloud

The wordcloud plot shows the popular words that appear in the document.

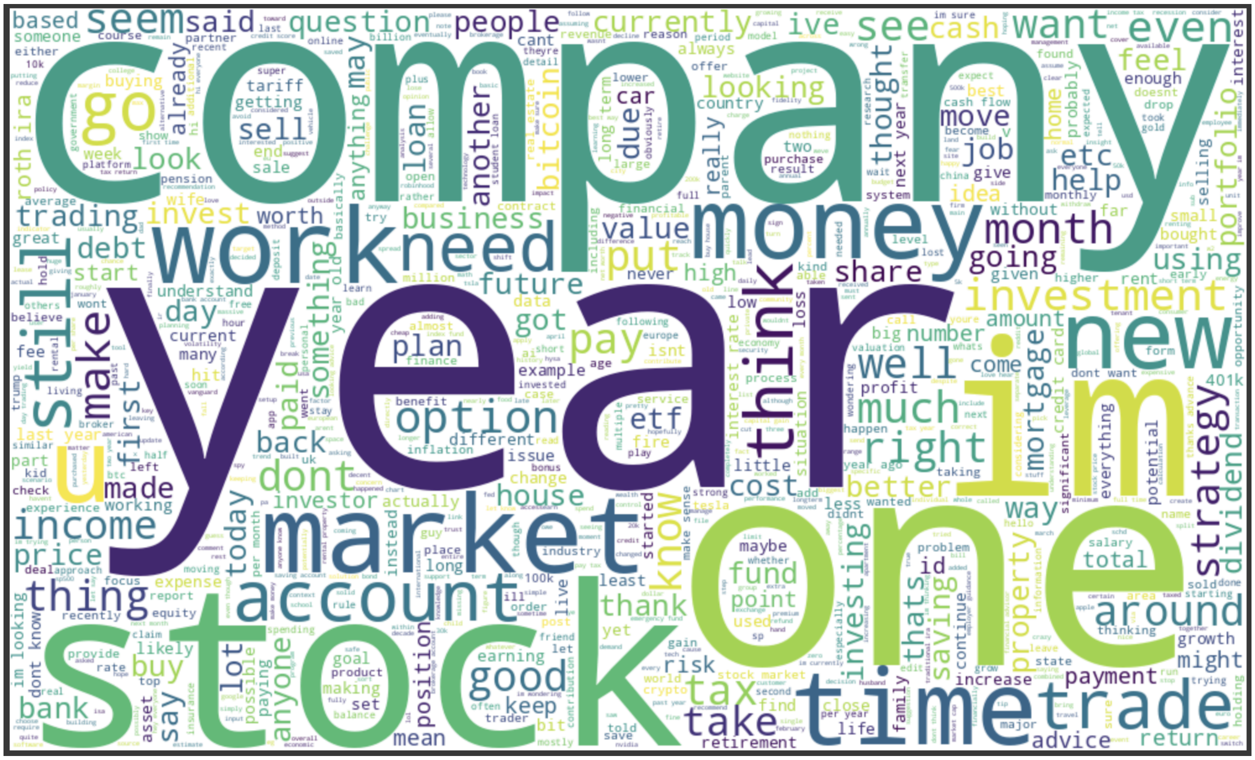


Figure 3

1. **Most Common Words:**
    - **"year":** Appears very large on the graph, suggesting frequent discussion of annual considerations, which could be yearly returns, year-end planning, or time horizons for investments.
    - **"company":** Likely used in the context of evaluating stocks, earnings reports, or business fundamentals**.**
    - **"one":** A generic term, but often appears in discussions comparing different options or scenarios (e.g., "one option," "one way," etc.).
    - **"stock"** and **"market"**: Their prominence indicates a strong focus on equities, stock market behaviour, or investing strategies.
    - **"money"**: A broad financial term, signifying discussions about capital, savings, and monetary decisions.
    - **“work”:** This term could reference employment, working hours, working in finance or whether investment strategies work.
    - **"Im"**: Not clear about its meaning, but it could likely be short for "I’m" (punctuations were removed during preprocessing), reflecting personal statements, opinions, or experiences shared in the text.
2. **Key Finance or Investing Terminology:**
    - **"stock", "investment", "etf", "funds", "dividend", "option", "trade", “crypto”, “bitcoin”**: These words point to various investing instruments (funds,

options), income sources (dividends), trading the financial market (trading), crypto currency trading like bitcoin and general financial management aspects.
    - **"mortgage"**: Suggests real estate financing topics, like purchasing a home or mortgage advice.
3. **Personal Finance Topics:** Words like **"pay," "people," "need," "work," "looking," "plus," "tax," "saving," "paid," "loan"** indicate personal finance concerns such as paying bills, dealing with taxes, taking or paying off loans, and saving money.
4. **Planning & Strategy:** Words like **“plan”**, “**strategy”**, **“goal”**, “**future”**, “**risk”**, and “**return”** indicate the intent to manage funds or assets and make strategic financial decisions, while assessing the risk.
5. **Time Orientation**: Words like **“year”, “month”, “day”, “time”, “today”, “next”,** and **“future”** suggest frequent temporal framing, important for investments and budgeting.
6. **Context and Themes:** The overall emphasis is on investments, market conditions, and personal financial planning.

MODEL IMPLEMENTATION AND FINE-TUNING

For this task, I will build two models, Latent Dirichlet Allocation (LDA) and BertTopic.

Common Hyperparameters

These are the hyperparameters that both BERTopic and LDA use.

- **random_state:** This seed value ensures reproducible results across different runs.
- **Number of topics (n_components for LDA or n_cluster for BERTopic):** This parameter determines the total number of topics or clusters into which the documents will be grouped. Choosing the right value is crucial. Too few topics can lead to very broad classifications, and too many topics can lead to highly fragmented classifications that are too specific and hard to interpret. Set to 10.

Latent Dirichlet Allocation (LDA):

The model building for Latent Dirichlet Allocation (LDA) is straightforward. Text representation is done using CountVectorizer, with the following hyperparameters.

- **Maximum document frequency (max_df):** Set to 0.9, meaning words appearing in more than 90% of the documents will be excluded. Words that are common in most documents won’t help the algorithm distinguish between topics.

- **Minimum document frequency(min_df):** This parameter excludes words that appear in fewer than the specified number of documents. It removes rare words that might add noise or cause overfitting to the topics. This parameter is set to 10.
- **Max_features:** Set to 10000**,** it limits the number of words to be considered during training. It helps to focus the analysis on the most relevant words and reduces computational complexity. Careful consideration should be taken when choosing this parameter, in order not to lose too much information.

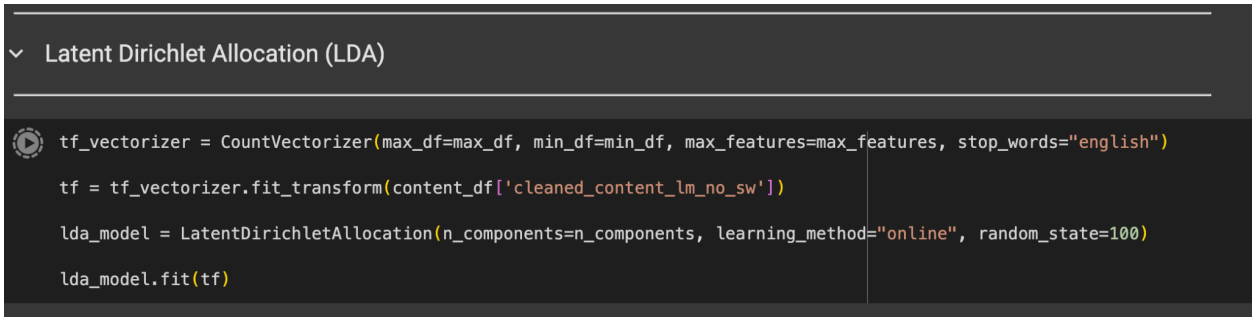


Figure 4

BERTopic

K-means: The model uses K-means instead of the default HDBSCAN with a cluster of 10, indicating 10 topics. I used K-means instead of HDBSCAN because I wanted to dictate the number of topics in the model and not let HDBSCAN figure out the number of topics automatically.

UMAP: Uniform Manifold Approximation and Projection is a popular dimensionality reduction technique.

Embeddings: The model uses **SentenceTransformer** and the **"all-MiniLM-L6-v2"** pretrained model, which is lightweight and very good when speed is desired.

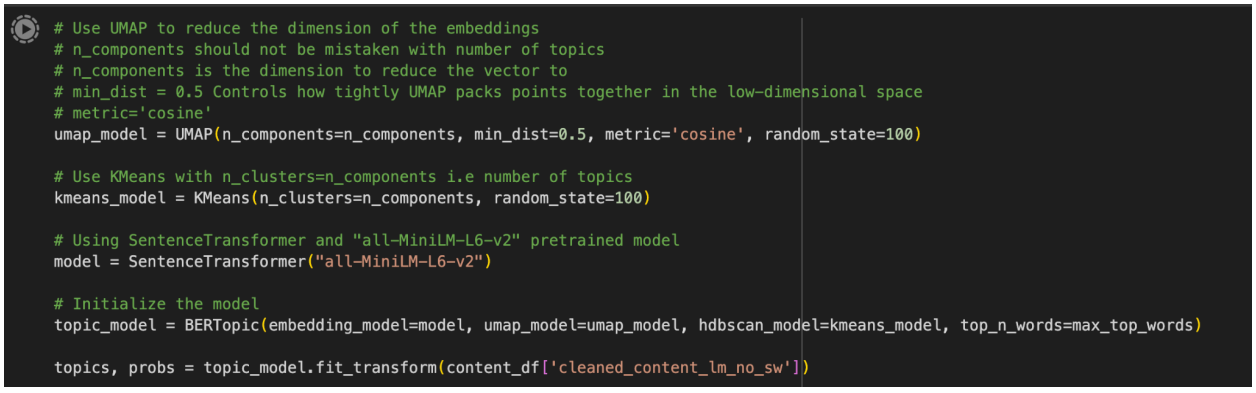


Figure 5

RESULTS AND VISUALISATIONS

The following visualisations show how the topics were made, they include bar graphs to show top words, and a Wordcloud that shows the most prominent words per topic. Dendrogram diagrams that show the relationship between topics and similarity matrices that show the similarity between topics.

Topic extracted using LDA

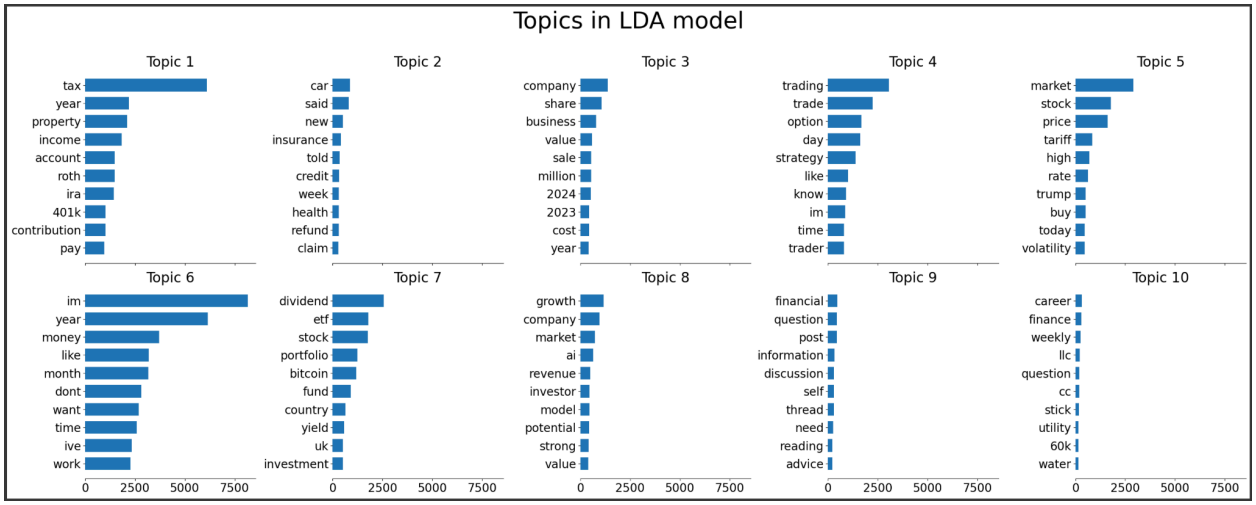
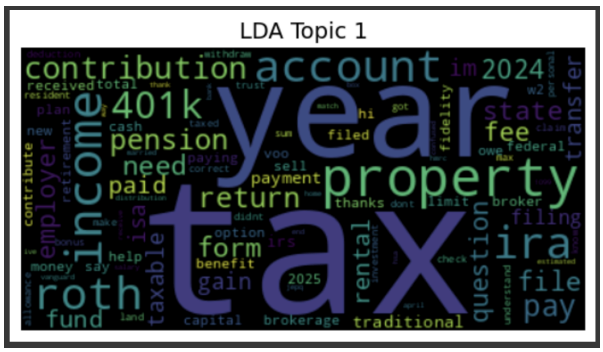
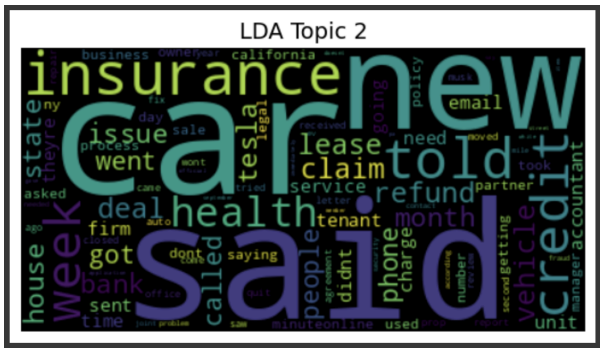
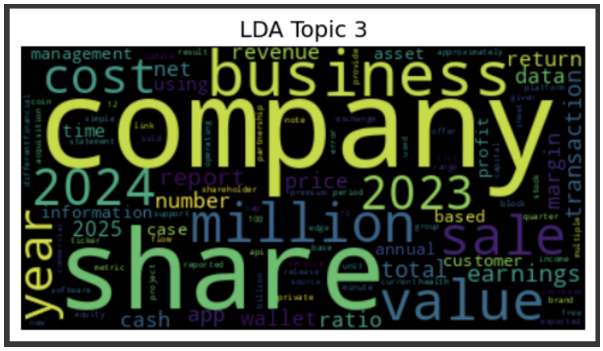
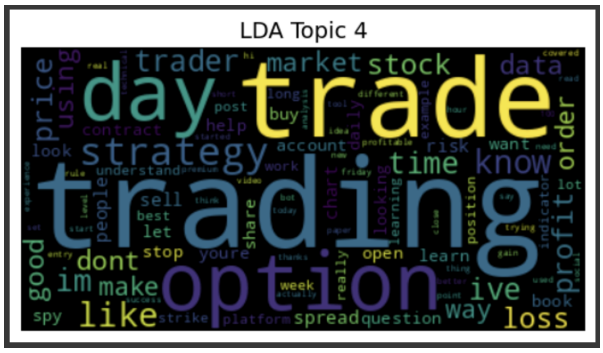
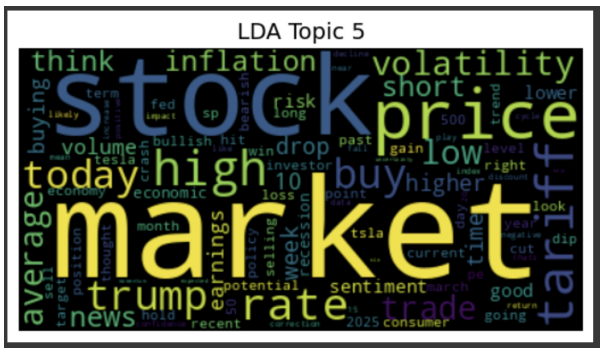
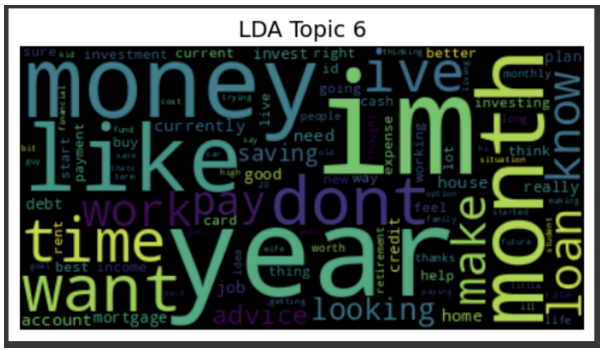
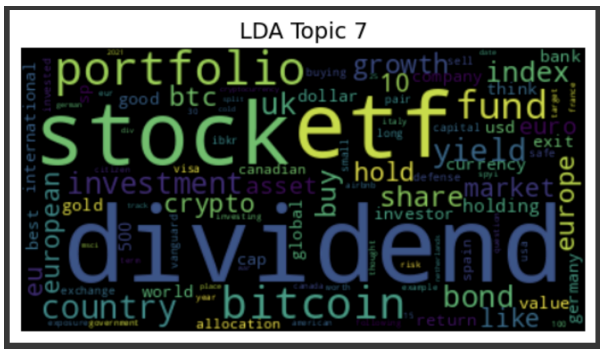
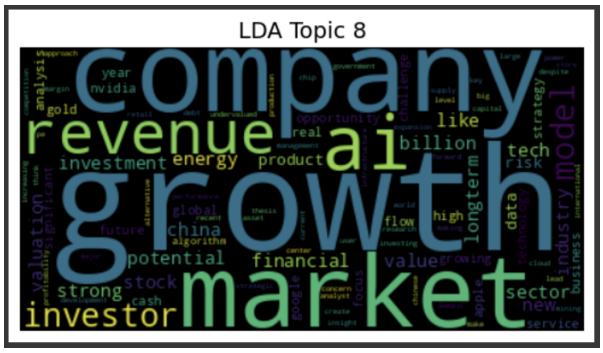
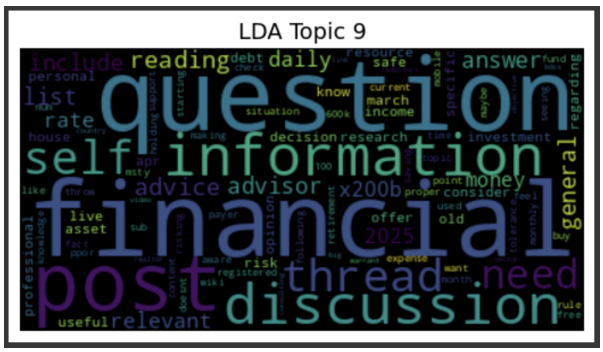
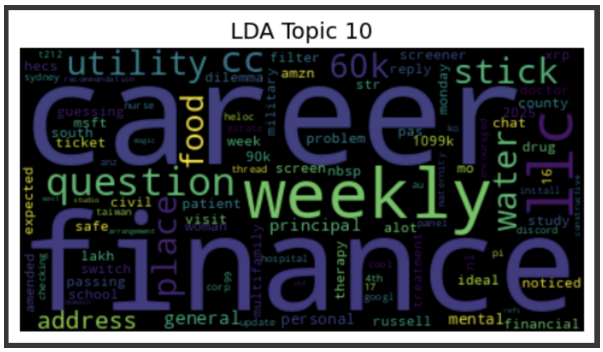


Figure 6

TOPIC	WORDCLOUD

The largest and most centrally located words are "tax", "year", "property," "return", "account", and "401k". These words suggest that Topic 1 is centred around finance, retirement and tax matters, likely concerning investments and retirement accounts.	 *Figure 7*
The most frequent and central words are "insurance", "car", "new", "said", "week", "issue", "health", "told", "claim", "deal" and "lease". These words suggest that Topic 2 is centred around insurance matters, particularly related to vehicles (cars) and potentially health.	 *Figure 8*
The largest and most central words are "company", "share", "business", "year", "million", "value", "cost", "sale", and the years "2024", "2023", and "2025". These words indicate that Topic 3 is focused on financial performance, business operations, and company valuation over a specific timeframe.	 *Figure 9*
The largest and most central words are "trade", "trading", "day", “option”, “strategy”, “time”, "stock", "market", "option", "time". These words indicate that Topic 4 is heavily focused on the activity of day trading in the stock market, also involving options trading.	 *Figure 10*

The largest and most central words are “market”, "stock", "price", "high", "buy", "rate", "trump", and "tariff". These words indicate that Topic 5 is focused on the stock market, specifically discussing stock prices, buying activity, interest rates, and the potential influence of Donald Trump and tariffs.	 *Figure 11*
The largest and most central words are "year", "money", "like", "work", "want", "month", "know", "don't", "I'm" and "live". These words indicate that Topic 6 revolves around personal financial situations, desires, and current circumstances, often expressed in a first-person perspective.	 *Figure 12*
The largest and most central words are "stock", "etf", "fund", "dividend", "investment", "portfolio", "market", "bitcoin", "crypto", "share" and "asset". These words indicate that Topic 7 is focused on investment strategies like stocks, Exchange Traded Funds (ETFs), mutual funds, dividend-paying assets, and cryptocurrencies like Bitcoin.	 *Figure 13*
The largest and most central words are "company", "growth", "market", "revenue", "ai", "investor", and "financial". These words indicate that Topic 8 is centred around the growth and financial performance of companies within the financial market, with emphasis on Artificial Intelligence (AI). This could be talking about the growth of AI companies or the use of AI by companies that foster growth.	 *Figure 14*

The largest and most central words are "question", "information", "financial", "post", "discussion", "self", "need", "advice" and "general". These words indicate that Topic 9 revolves around seeking and sharing financial information, asking questions, and engaging in general discussions about personal finance.	 *Figure 15*
The largest and most central words are “career”, “weekly”, ‘finance”, “question”, “safe”, “utility”, “cc”, “stick”, and “water”. They suggest that Topic 10 revolves around career, employment, financial security, social security, water and utility bills.	 *Figure 16*

Topic extracted using BERTopic

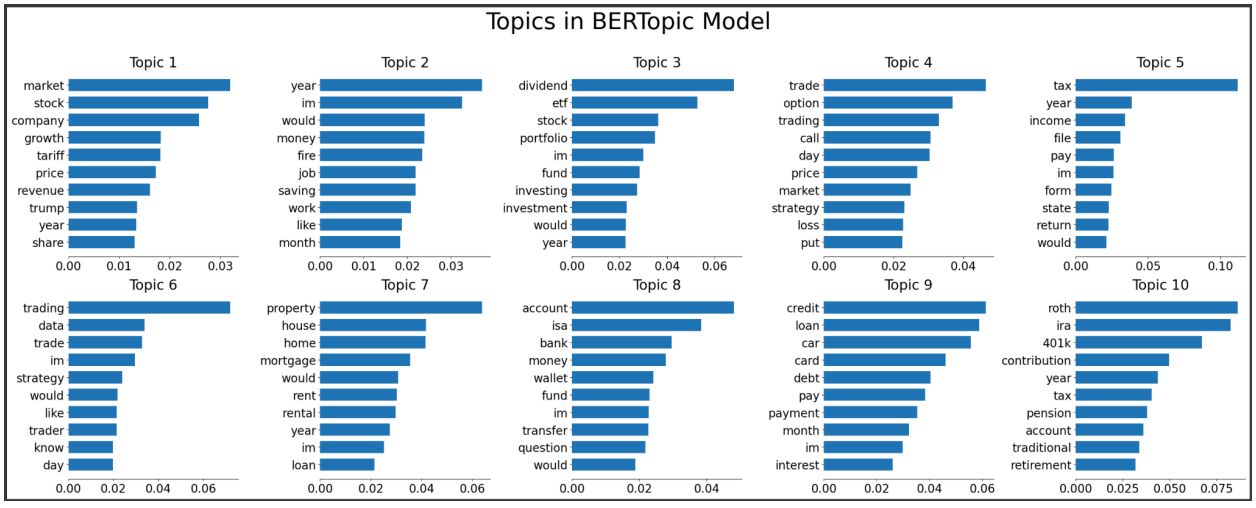
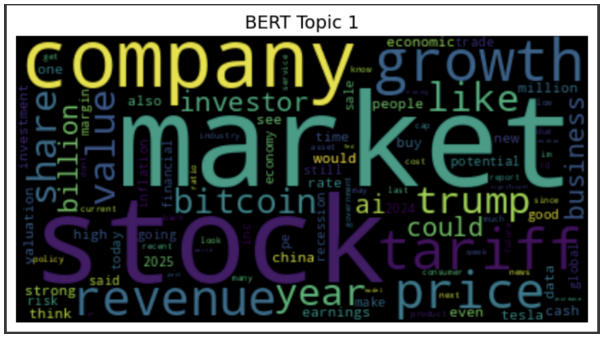
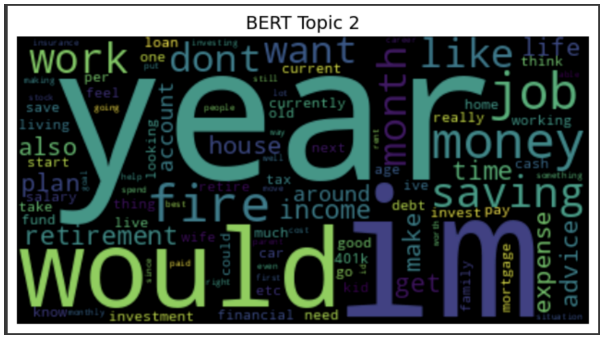
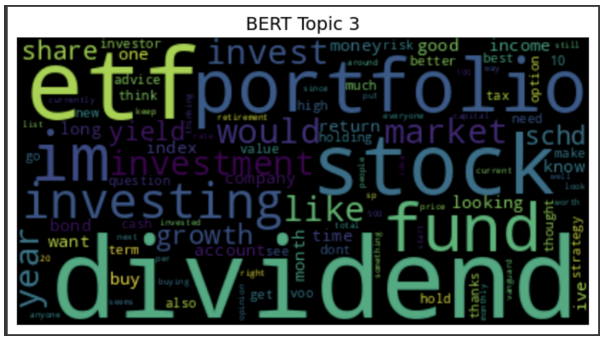
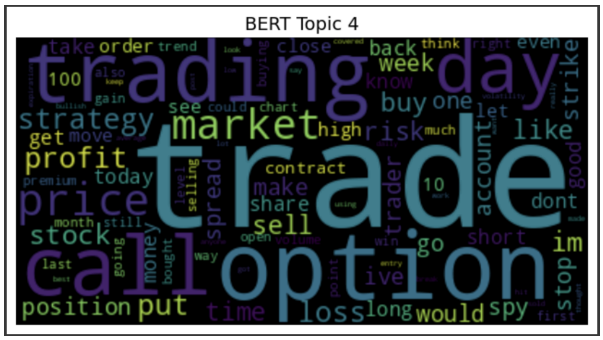
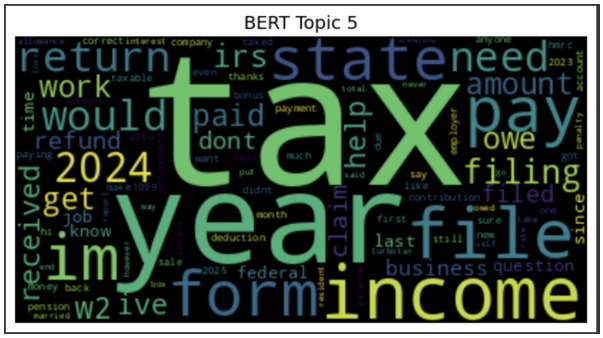
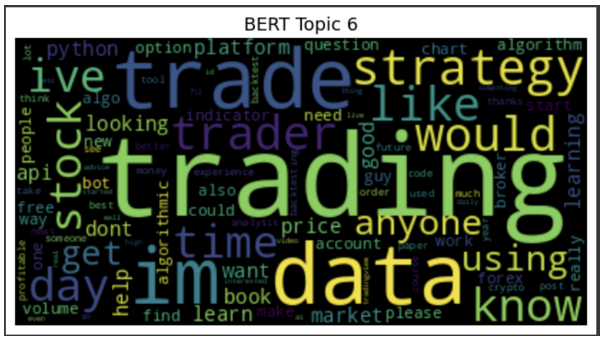
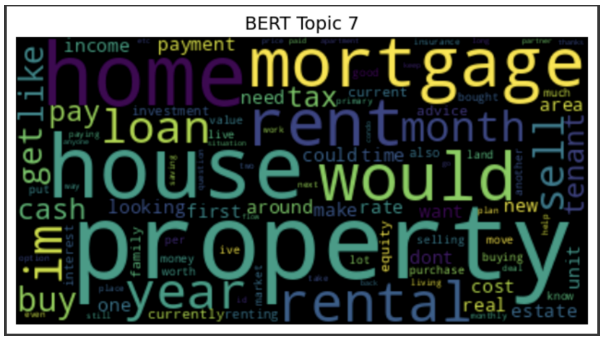
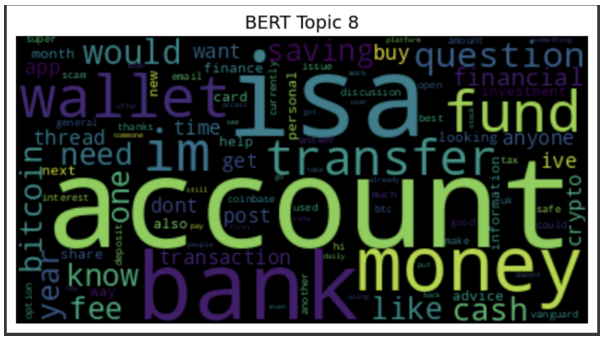
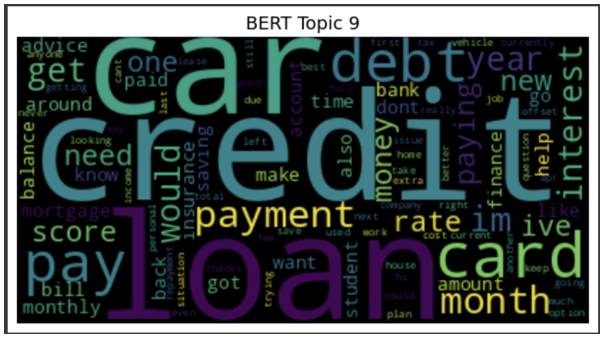
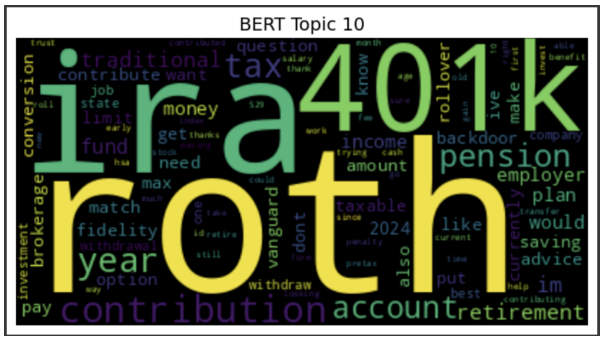


Figure 17

TOPIC	WORDCLOUD

The largest and most prominent words are "market", "stock", "company", "revenue", "tariff", "growth", "price", "share", "trump", and "value". These words indicate that Topic 1 revolves around economic activity, business performance, investment and the impact of Donald Trump with tariffs.	 *Figure 18*
The largest and most prominent words are "year", "work", "job", "money", "saving", "retirement", and "fire" (likely referring to Financial Independence, Retire Early). These words indicate that Topic 2 is centred around personal finance, work-life, future planning, and retirement.	 *Figure 19*
The largest and most prominent words are "dividend", "etf", "portfolio", "stock", "investing", "investment", "investing", "fund", and "share". These words indicate that Topic 3 revolves around investments across different asset classes and approaches.	 *Figure 20*
The largest and most prominent words are "trade", "trading", "option", "call", "put", “sell”, “position”, "price" and "day". These words indicate that Topic 4 centres around buying and selling options contracts.	 *Figure 21*

The largest and most prominent words are "tax", "year", "file", "filing", "return", "form", "income", "pay", "irs", and "state". These words indicate that Topic 5 centres around tax, income tax, tax preparation, and tax filing. This topic is all about taxation and income tax.	 *Figure 22*
The largest and most prominent words are "trading", "trade", "strategy", "algorithm", "data", "trader", "algorithmic", "stock", "algo", "bot", "python", "api", and "code". These words indicate that Topic 6 is focused on algorithmic and data-driven trading strategies.	 *Figure 23*
The largest and most prominent words are "home", "mortgage", "house", "property", "rent", "loan", "buy", "selling", "purchase" and "cost". These words indicate that Topic 7 is focused on properties, housing, house rent, residential living and mortgage.	 *Figure 24*
The largest and most prominent words are "account", "bank", "money", "transfer", "wallet", "card", “fee”, "fund", "bitcoin" and "crypto". These words indicate that Topic 8 is focused on managing and transferring money (including cryptocurrencies) online.	 *Figure 25*

The largest and most prominent words are "car", "debt", "credit", "loan", “pay”, "payment", "interest", "rate", "balance", "score", and "bill". These words indicate that Topic 9 is centred around debt management and borrowing, particularly related to car ownership through loans.	 *Figure 26*
The largest and most prominent words are "401k", "ira", "roth", "pension" and "retirement". These words indicate that Topic 10 is centred around retirement savings plans, specifically highlighting 401k and IRA (Individual Retirement Account) accounts.	 *Figure 27*

Hierarchical clustering dendrogram diagram

The dendrogram shows the relationship between topics. It uses branches like a tree, and the distance between branches shows whether the topics are similar or different. Topics that are similar appear closer on the dendrogram, whereas topics that are not similar to each other appear farther.

LDA:

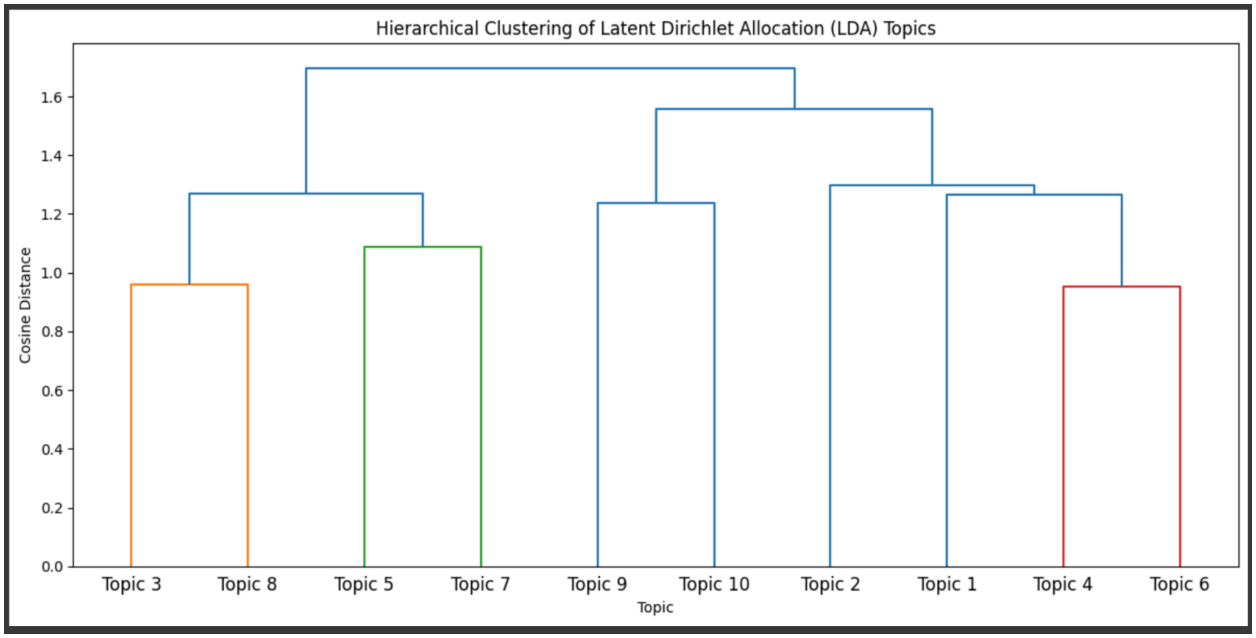


Figure 28

The above dendrogram shows that Topics 3 and 8 are connected at a lower (or about) 0.95 distance, meaning they are similar. At a somewhat lower distance (about 1.1), there are clusters formed by Topics 5 and 7. At a distance of around 1.24, topics 9 and 10 are grouped. At a distance of about 1.3, topics 2 and 1 are linked. At a distance of about 0.95, topics 4 and 6 are joined, signifying similarity.

Moving up the dendrogram, the initial clusters are being merged. The cluster of Topics 3 and 8 is joined with Topics 5 and 7 at a higher distance (about 1.27), indicating a less direct relationship between these two groups compared to the relationships within each group. The overall structure of the dendrogram and the heights at which the main branches merge give an idea of how distinct the different sets of topics are. The final merging of the two main branches occurs at a relatively high distance (above 1.5), suggesting a considerable difference between the two major groups of topics identified.

BERTopic:

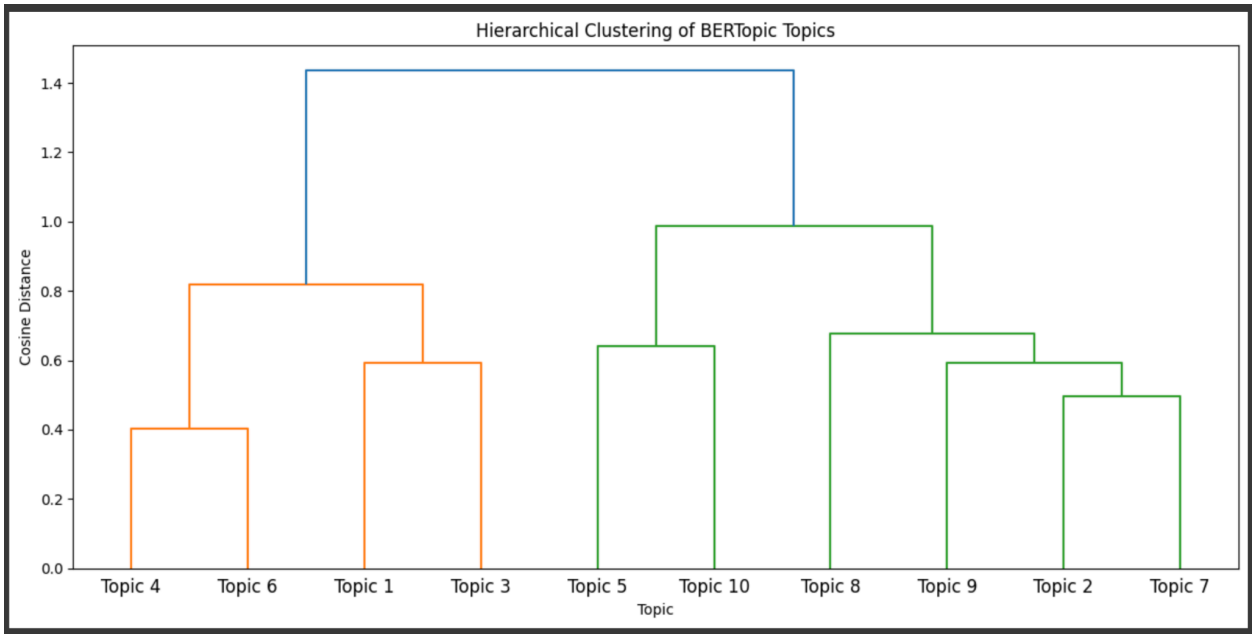


Figure 29

The plot above shows that, Topics 4 and 6 are joined at a very low distance (about 0.4), indicating a high degree of semantic similarity. Topics 1 and 3 are clustered together at a distance of about 0.6, suggesting a notable semantic relationship. Topics 5 and 10 are joined at a distance of about 0.65, indicating semantic similarity. Topics 8 and 9 are clustered around a distance of 0.7, suggesting a relationship. Topics 2 and 7 are joined at a relatively low distance (about 0.5), indicating semantic closeness.

Moving up the dendrogram, these initial clusters are further merged, revealing broader semantic groupings. Topics 4 and 6 cluster is joined with the Topics 1 and 3 cluster at a higher distance (about 0.82), suggesting a less direct semantic link between them. Overall Semantic Distinctiveness: The height at which the main branches of the dendrogram merge (above 1.4) indicates a substantial semantic difference between the major groups of topics identified by

Overall, both dendrograms have roughly the same shape, indicating that they grouped the documents into similar topics. However, Bertopic created topics that were more similar than LDA. Notice how the topics within the Bertopic dendrogram have closer relationships, while LDA topics are more distinct.

Similarity Matrix

The Similarity Matrix shows the relationship between topics, on a scale of zero to one, where values closer to zero mean low similarity, while values closer to one mean high similarity.

LDA:

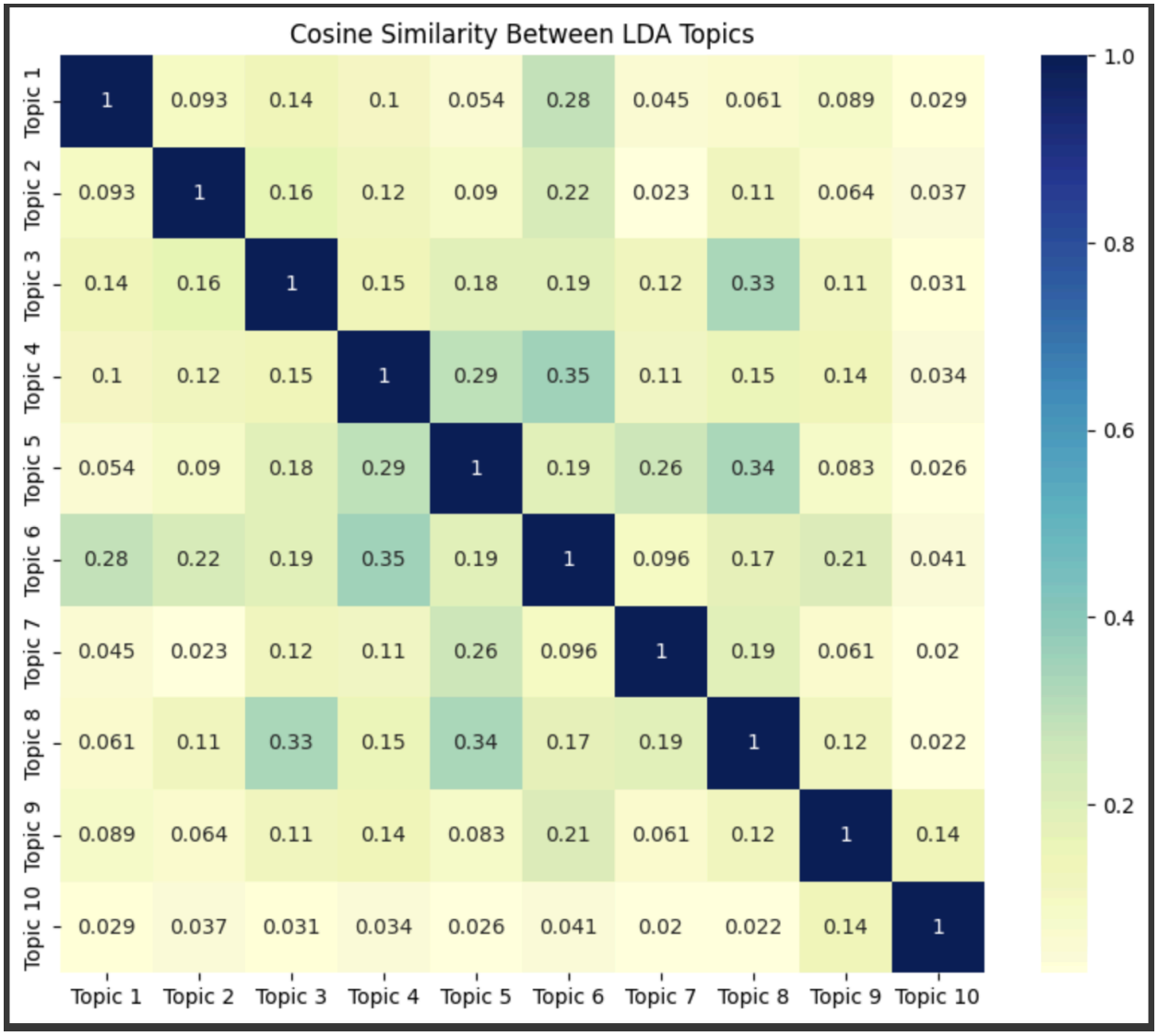


Figure 30

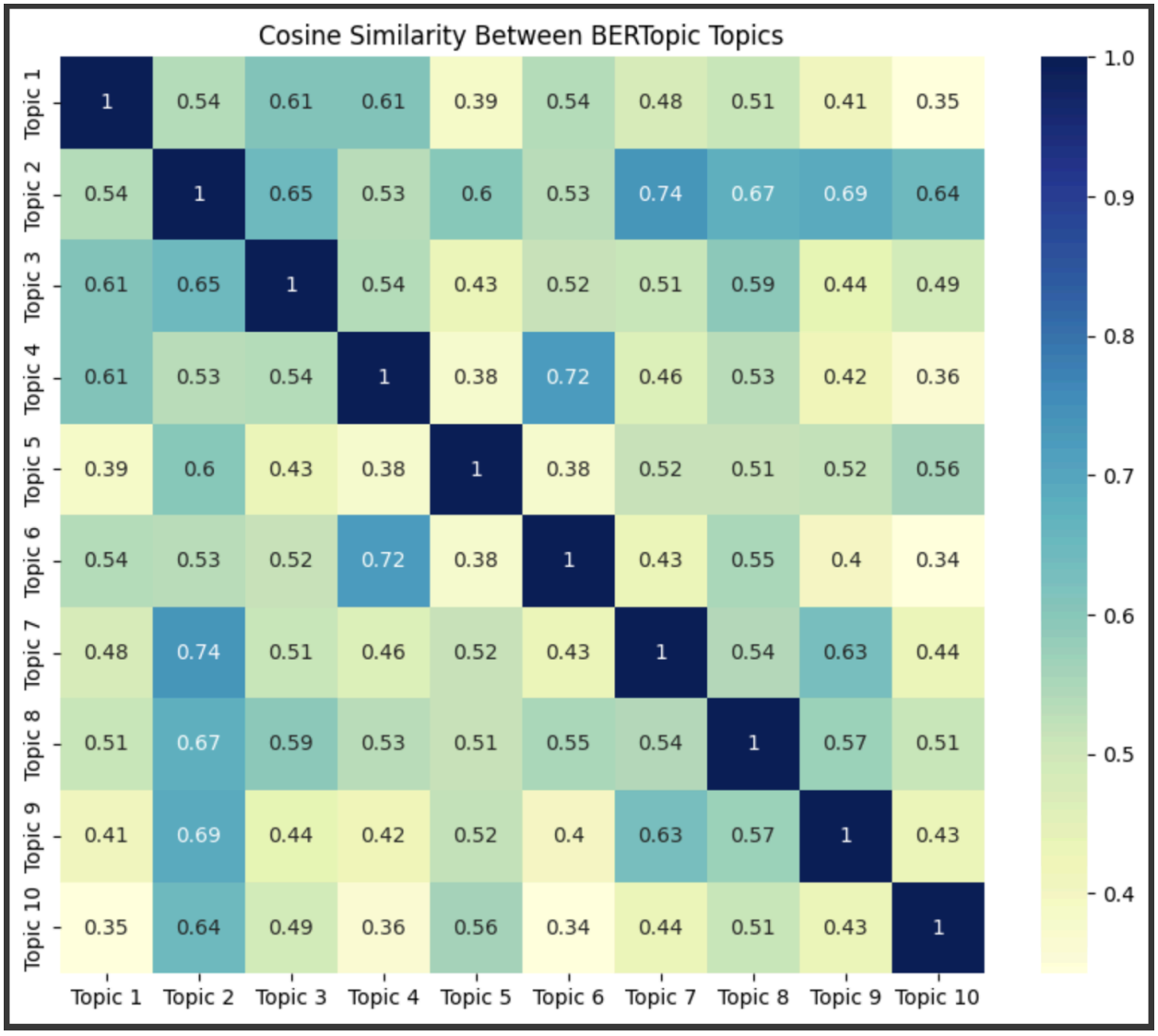


Figure 31

COMPARISON BETWEEN GOOGLE COLAB AND LINUX

Attribute	**Google Colab (cloud T4)**	**Linux machine (local 32 GB RAM)**
The libraries used in building the models and their versions	Most of the libraries used were already installed on Colab, so no need to install them again. However, the latest versions of Bertopic and Praw were installed:	The following were installed: “jupyterlab==4.4.0”, “notebook==7.4.0”, “numpy==2.2.5”, “pandas==2.2.3”, “seaborn==0.13.2”,

	“bertopic==0.16.4” and “praw==7.8.1”.	“matplotlib==3.10.1”, “nltk==3.9.1”, “wordcloud==1.9.4”, “praw==7.8.1” and “bertopic==0.16.4”.
Speed of training LDA	It took about 33 seconds to train.	It took about 17 seconds to train.
Speed of training Bertopic (Excluding the time it took to download the embedding)	It took 60 seconds to train.	It took about 19 seconds to train.

Conclusion

In this paper, I explored financial discussions on Reddit using two topic modelling techniques, Latent Dirichlet Allocation (LDA) and BERTopic. Both methods were very good at classifying the topics and categorising them similarly, as evident in the dendrograms having the same shape. However, Latent Dirichlet Allocation (LDA) categorises topics more distinctly than BERTopic. This is evident in the dendrograms and the similarity matrix, with LDA topics having the highest similarity of 0.35 and the lowest similarity of 0.02. On the other hand, BERTopic had the highest similarity of 0.74 and the lowest similarity of 0.34.

Regarding speed and performance, Latent Dirichlet Allocation (LDA) was about two times faster than Bertopic, and is an excellent choice where computational resources are low and speed is desired. However, some topics generated by Latent Dirichlet Allocation (LDA) didn’t make sense in a financial context, particularly Topic 9, which has no financial bearings. This indicates that Bertopic is more context-aware and LDA just groups words based on probabilities.

REFERENCE

Shivam Bansal (2025, Feb 17). What is Topic Modeling?. Retrieved from https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python