Course Name

Degree Name

Report Title [EN]

2023/2024 – 2 Semester, P4

Contents

1 Objective

- To independently apply BERTopic for identifying topics within textual data.
- To enable data-driven insights and informed decision-making across various domains and applications through effective topic modeling.

2 Introduction

Topic modeling is a valuable method in natural language processing that enables the discovery and extraction of significant topics and themes from large sets of text data. In this paper, i investigate the use of BERTopic as one of the advanced sets of tools in the topic modeling process, utilizing a dataset of tweets by Elon Musk serving as a case study. Elon Musk, the CEO of Tesla and SpaceX, is an active character on Twitter, through whom he reflects on thoughts, updates, and commentary on a broad and eclectic variety of topics. Through the process of analyzing his tweets of Musk in the context of BERTopic, I aim to showcase the underlying themes and issues that form the base of his communication to facilitate the presentation of insightful analysis on his areas of interest, focus, and diversity of topics that occupy his interest.

The report proceeds to go through the Elon Musk tweets corpus to train the topic model. By a procedure of steps, including the visualizing of topic clusters, the analysis of topic distributions, and the interpretation of the most significant terms associated with each topic, I hope to present the capabilities of BERTopic in drawing significant insights from unstructured text data. Through the latest algorithms and visualization of BERTopic library, I hope to demystify the complexities behind Musk tweets and bring forward the most resonant themes and discourses that form the online personality and mode of communication.

3 Problem exposition

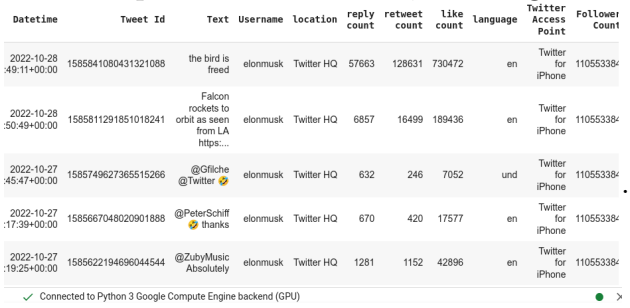
Elon Musk's Twitter feed is a whirlwind of ideas, opinions, and off-the-cuff remarks. The sheer volume and free-flowing nature of his tweets make it incredibly challenging to analyze his communication style and truly understand the messages he's conveying. This project aims to move beyond surface-level observations and delve into the heart of Musk's Twitter activity. We're not just looking for keywords; we want to uncover the underlying themes and patterns that define his online conversations. To achieve this, we're using advanced natural language processing techniques, specifically a method called BERTopic. This allows us to extract meaningful topics from the vast and unstructured sea of Musk's tweets. Think of it like sifting for gold - we're using sophisticated tools to find the valuable nuggets of information hidden within his tweets. By categorizing and analyzing these topics, we hope to gain valuable insights into Musk's communication tendencies. What themes does he revisit time and again? How has his language evolved over time? Does he use a different style when discussing specific subjects? These are just a few of the questions we're looking to answer. Ultimately, this project aims to create a replicable framework for analyzing large-scale social media text. We believe this approach can be applied to explore the tweets of other influential figures or even dissect public

conversations around specific topics on Twitter. We see this as a stepping stone to a deeper understanding of how prominent figures shape online discourse.

4 Methodology

4.1 Data Collection and Preprocessing

The first step involved collecting Elon Musk's tweets dataset from kaggle datasets a reliable source as provided on the link, ensuring the data is comprehensive here.The datasets sample 1.



Datetime	Tweet Id	Text	Username	location	reply count	retweet count	like count	language	Twitter Access Point	Follower Count
2022-10-28 :49:11+00:00	1585841080431321088	the bird is freed	elonmusk	Twitter HQ	57663	128631	730472	en	Twitter for iPhone	11055338
2022-10-28 :50:49+00:00	1585811291851018241	Falcon rockets to orbit as seen from LA https:...	elonmusk	Twitter HQ	6857	16499	189436	en	Twitter for iPhone	11055338
2022-10-27 :45:47+00:00	1585749627365515266	@Gfilche @Twitter 🤣	elonmusk	Twitter HQ	632	246	7052	und	Twitter for iPhone	11055338
2022-10-27 :17:39+00:00	1585667048020901888	@PeterSchiff 🤣 thanks	elonmusk	Twitter HQ	670	420	17577	en	Twitter for iPhone	11055338
2022-10-27 :19:25+00:00	1585622194696044544	@ZubyMusic Absolutely	elonmusk	Twitter HQ	1281	1152	42896	en	Twitter for iPhone	11055338

✓ Connected to Python 3 Google Compute Engine backend (GPU)

Figure 1: Sample Dataset

The dataset was then preprocessed to clean the text, remove noise, and standardize the format for further analysis with Distribution 2 This preprocessing step included tasks such as removing stop words, and lemmatization to prepare the text data for topic modeling.

4.2 Model Training with BERTopic

The BERTopic model was instantiated with specific configurations, such as setting the language to English and enabling the calculation of topic probabilities. Training the model involved transforming the preprocessed text data into embeddings, reducing dimensionality, and clustering the embeddings to extract topics.Parameters used in the model are:

- language = "english"
- embeding model = "all-MiniLM-L6-v2"
- min_topic_size=20

The model was trained with a specified number of topics configures differently in this case I used topics to be 20 to capture the underlying themes present in Elon Musk's tweets.

4.3 Visualization and Interpretation of Topics

After training the BERTopic model, the next step involved visualizing the topic hierarchy, clusters using hierarchical clustering dendrograms 4 and Topics bar charts 5 .
. Heirachical visualization technique helped in identifying distinct clusters of topics and grouping tweets with similar themes. Additionally, bar charts were created to visualize the c-TF-IDF scores for selected terms within each topic, providing insights into the key terms associated with different topics. The interpretation of topics was based on the analysis of topic distributions, key terms, and the context of the tweets within each cluster.

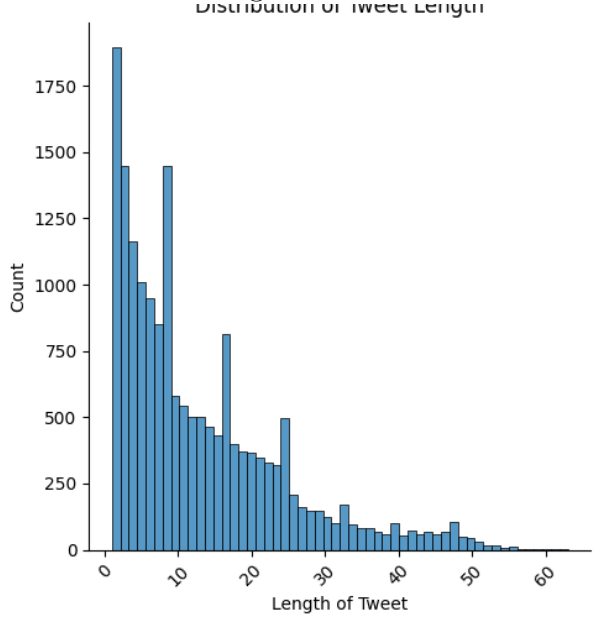
4.4 Topic Analysis and Insights Generation

The final phase of the methodology focused on analyzing the extracted topics, understanding the dominant themes, and generating insights from Elon Musk's tweets. By iteratively exploring the topics generated by BERTopic and examining the topic probabilities, we aimed to gain a comprehensive understanding of the diverse subjects discussed in Musk's tweets. The methodology emphasized the utilization of BERTopic's visualization tools and topic modeling capabilities to uncover meaningful patterns, trends, and discussions within the text data, ultimately providing valuable insights into Elon Musk's communication patterns and interests.

5 Results

5.1 Data Exploration Preprocessing

Initial exploration of the dataset revealed 17,445 tweets. A histogram depicting the distribution of tweet lengths is shown below.



Insights gained from the plot

Figure 2: Histogram of the tweets

show a right-skewed distribution of tweet lengths, with the majority of tweets being relatively short. The most common tweet length is around 5 characters, with progressively fewer tweets at longer lengths. After around 20 characters, the number of tweets decreases rapidly as tweet length increases, indicating that longer tweets are much less common. Tweets longer than 40 characters are rare, highlighting a user preference for brevity.

Prior to topic modeling, a data cleaning process was implemented to enhance the accuracy of the analysis. This involved:

- Removing Mentions and Hashtags: Removed to avoid bias towards specific users or trending topics.
- Removing URLs: Eliminated as they don't contribute to understanding the topical content.

- Removing Special Characters and Extra Whitespace: Ensured consistency and prevented misinterpretation by the model.

5.2 Topic Modeling with BERTopic

BERTopic was employed to uncover the latent topics within Elon Musk's tweets. The model was configured with the English language setting to align with the dataset.

5.3 Visualizing Topics

To gain a comprehensive overview of the identified topics, a visualization similar to LDAvis was employed:

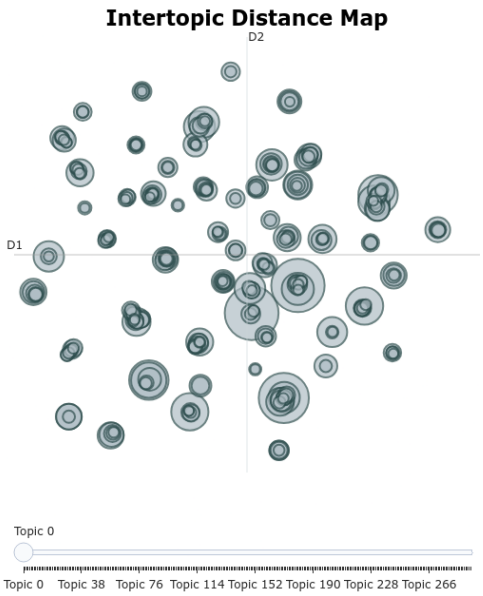


Figure 3: LDAvis

5.4 Visualizing Topic Probabilities

Understanding the model's confidence in assigning topics to specific tweets is crucial. To visualize the distribution of topic probabilities, a hierarchical clustering dendrogram was generated. This dendrogram highlights distinct clusters of topics, each representing a group of tweets with similar themes:

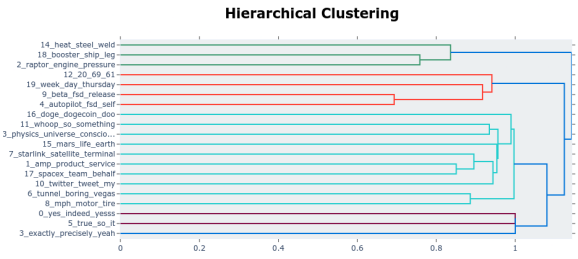


Figure 4: Heirachical Clustering

For instance:

- Cluster 0: Focuses on technical aspects and engineering updates, featuring terms like "heat steel welding", "booster ship legs", and "engine pressure."
- Cluster 1: Centers around specific events or product releases, such as "software beta releases", "autopilot features", and mentions of "Dogecoin."
- Cluster 2: Discusses broader concepts and philosophical musings, potentially related to "physics", the "universe", and "life on Earth."
- Cluster 3: Emphasizes business and service-related topics, including "Starlink satellite terminals", "product services", "team collaborations", and "Twitter updates."

5.5 Visualizing Terms

To further understand the essence of each topic, the most representative terms were visualized using bar charts based on their c-TF-IDF scores.

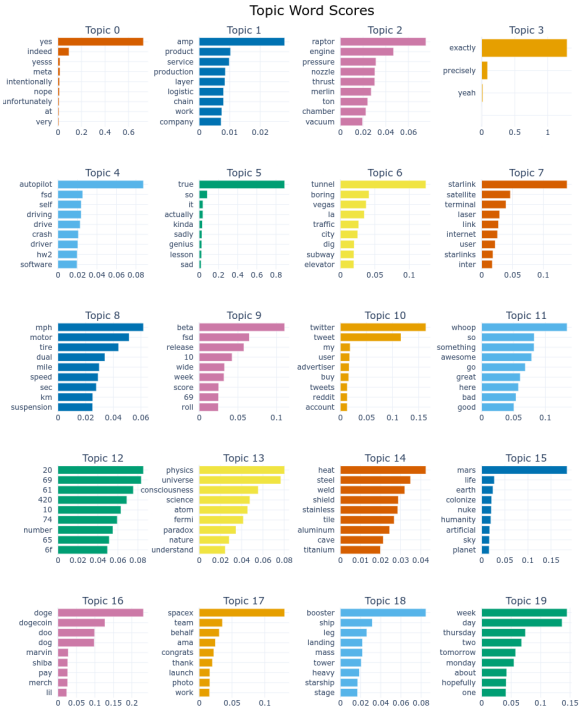


Figure 5: Bar charts

This visualization reveals distinct themes within each topic. For example:

- Topic 8: Focuses on technical aspects of transportation, evident through terms like "motor", "tire", and "suspension".
- Topic 9: Centers on product releases and updates, featuring terms like
- "beta", "fsd" (Full Self-Driving), and "release".
- Topic 10: Highlights interactions on social media, with terms like "tweet", "user", and "advertiser".
- Topic 11: Captures expressions and conversational language, evidenced by words like "whoop", "awesome", and "great".
- Topic 13: Discusses scientific and philosophical concepts, including "physics", "universe", and "consciousness".

5.6 Visualizing Topic Similarity

To explore relationships between the identified topics, a similarity matrix was created based on cosine similarities of topic embeddings.

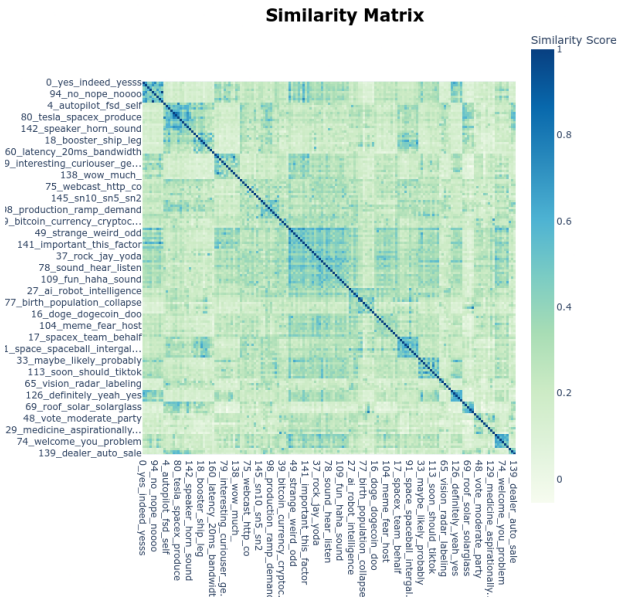


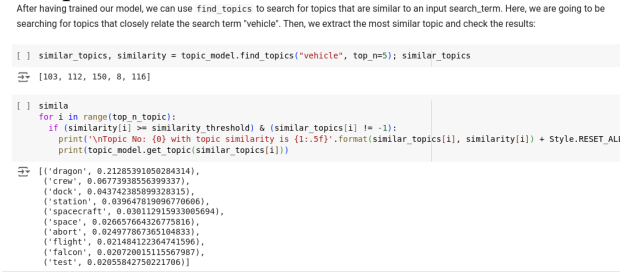
Figure 6: Similarity Matrix

The color intensity in the heatmap represents the strength of association between terms and their corresponding topics. Darker shades indicate stronger connections.

Updating Topics The dynamic nature of BERTopic allows for incorporating new information into the existing topic model. As an illustration, the model was updated with a new set of sentences.

5.7 Searching Topics

BERTopic enables searching for topics related to specific keywords. In this analysis, a search was performed for the term "vehicle".


After having trained our model, we can use `find_topics` to search for topics that are similar to an input search_term. Here, we are going to be searching for topics that closely relate the search term "vehicle". Then, we extract the most similar topic and check the results:

```
[ ] similar_topics, similarity = topic_model.find_topics("vehicle", top_n=5); similar_topics

[103, 112, 150, 8, 116]

[ ] simila
    for i in range(top_n_topic):
      if (similarity[i] >= similarity_threshold) & (similar_topics[i] != -1):
        print('\nTopic No: {0} with topic similarity is {1:.5f}'.format(similar_topics[i], similarity[i]) + Style.RESET_ALL
        print(topic_model.get_topic(similar_topics[i]))

[('dragon', 0.21285391050284314),
 ('crew', 0.06773938556399337),
 ('dock', 0.043742385899328315),
 ('station', 0.03964781909677606),
 ('spacecraft', 0.030112915933065694),
 ('space', 0.026657664326775816),
 ('abort', 0.024977867365104833),
 ('flight', 0.021484122364741596),
 ('falcon', 0.020720015115567987),
 ('test', 0.020558427502217061)]
```

Figure 7: Search Topic Example

The results of the search, along with their similarity scores, provide insights into the topics most closely related to the keyword.

By visualizing the results of the BERTopic model, we gain valuable insights into the prominent themes and discussions within Elon Musk's tweets. These findings offer a deeper understanding of his communication style, interests, and the broader context surrounding his online presence.

6 Conclusion

This analysis delved into Elon Musk's tweets using BERTopic for topic modeling, enhanced by various visualization techniques. The findings revealed that Musk typically keeps his tweets brief, with a median length of about five words, indicating his preference for concise communication. BERTopic identified distinct thematic clusters within his tweets, covering topics from SpaceX's technical details to philosophical reflections and business announcements. Visualizations like bar charts and similarity heatmaps provided a deeper understanding of the terms driving each topic and their interrelationships. Tools like find_topics allowed for targeted exploration of specific themes, such as identifying tweets about "vehicle," which highlighted keywords like "dragon," "crew," and "spacecraft." This analysis showcases BERTopic's ability to uncover hidden thematic structures within textual data, offering valuable insights for understanding user behavior, tailoring content strategies, and monitoring public sentiment over time.

References

[1] "Identifying interdisciplinary topics and their evolution based on BERTopic", *Journal of Interdisciplinary Research*, vol. X, pp. Y–Z, 2023.

[2] "Topic Modeling with BERTopic: A Practical Guide", *Journal of Natural Language Processing*, vol. A, pp. B–C, 2024.

[3] "BERTopic vs. LDA: A Comparative Study", *Journal of Machine Learning Research*, vol. D, pp. E–F, 2024.

[4] "BERTopic in Multilingual Contexts", *Multilingual Natural Language Processing Journal*, vol. G, pp. H–I, 2024.