Student Name:
Gunpreet Singh

Student Number:
1296550

Course Code: DATA*6700

Project Title:
The Globalizing K-pop Project: Sentiment Analysis of Global Fandoms on Social Media

Submission Date: 12th August 2024

Name of Primary Grader
Dr. Stacey Scott
Name of Second Grader
Dr. Ritu Chaturvedi

Table of Contents

1. ABSTRACT

The project is a comprehensive study aimed at understanding the role of social media in enhancing the well-being and sense of community among global K-pop fandoms. This research leverages sentiment analysis using topic modeling techniques on data extracted from Twitter and Reddit, focusing on various forms of social support, including emotional, informational, and appraisal, within these fandoms. The study employs topic models such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Non-Negative Matrix Factorization (NMF), and BERTopic, to analyze the collected data, which has been preprocessed through steps like tokenization, lemmatization, and keyword filtering. The research workflow involves extracting top keywords, visualizing clusters, and using representation models such as Llama2, and KeyBERT to represent the results to the stakeholders. The findings highlight key themes within the fandoms, such as emotional investment, mental health, and fan culture, underscoring the significance of social media in connecting diverse fans across different geographic regions and cultures. The project aims to improve access to Korean K-pop media, thereby promoting inclusivity and community well-being. The BERTopic model achieved satisfactory performance with a Cv coherence score of 0.59, which is commendable given the limited dataset available. Among the models tested, Latent Dirichlet Allocation (LDA) demonstrated the highest effectiveness, yielding a Cv coherence score of 0.63. The Non-negative Matrix Factorization (NMF) model produced a moderate score of 0.54, while the Latent Semantic Analysis (LSA) model exhibited the lowest coherence score of 0.36. These results highlight the relative strengths and limitations of each modeling approach within the context of the data used.

2. INTRODUCTION

2.1 Background

Sentiment analysis, an essential tool in understanding opinions and emotions expressed on social media, has significantly advanced with the integration of natural language processing (NLP) techniques and machine learning models. These advancements have been particularly effective in classifying text data into sentiment categories such as positive, negative, or neutral. Topic modeling, a method traditionally used to discover the underlying themes in a corpus of text, is increasingly being leveraged in sentiment analysis to provide a more granular understanding of sentiments across different topics within large datasets [1].

The advent of deep learning models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa has further enhanced sentiment analysis by capturing the contextual nuances of text. These models, when combined with topic modeling techniques such as BERTopic, allow for the identification of specific sentiment trends across various topics, making the analysis more targeted and insightful [1].

2.2 Problem Statement

Despite the progress in sentiment analysis, there is still a significant challenge in effectively capturing the nuanced sentiments expressed across multiple topics in social media data. Traditional sentiment analysis models often lack the ability to simultaneously address the complexity of sentiments and the diversity of topics present in large, multilingual, and contextually varied datasets [1]. Additionally, the integration of advanced topic modeling techniques like BERTopic with sentiment analysis has not been fully explored, especially in dynamic environments where both topics and sentiments evolve rapidly over time [1].

2.3 Gap in Knowledge

There is a noticeable gap in the current research that involves combining advanced sentiment analysis models (e.g., BERT, RoBERTa) with sophisticated topic modeling approaches like BERTopic to analyze social media data. Existing studies have not thoroughly investigated how these integrated models can be optimized to better understand and predict the sentiments associated with specific topics in multilingual and dynamic social media environments. Addressing this gap could significantly improve the accuracy and relevance of sentiment analysis in these complex contexts [1].

3. METHODS

3.1. Data Collection

3.1.1 Twitter Data: We got twitter data which was collected using twitter API in 2023. This data is related to K-pop Sf9 fandom.

3.1.2 Web Scraping: Using PRAW (Python Reddit API Wrapper) for web scraping.

Targeted K-Pop Music:

- Bangtan Boys (BTS)
- BlackPink
- SF9
- Kpop_uncensored
- K-pop

3.2. Data Preprocessing

3.2.1. Linking Conversations:

For SF9 data collected using the Twitter API, we have two main columns related to text: "first_tweet" and "text." The 'first_tweet' column is identical for data that is part of the same conversation, as it represents the first tweet of the conversation. The "text" column contains the replies within the conversation. To structure this data, we concatenated the conversation using the conversation ID, which is common across the conversation. Additionally, we prefixed the concatenated replies with the "first_tweet".

```
df.head()
```

	text	translated_text	cleaned_text
0	RM's 'Right Place, Wrong Person' Compilation G...	RM's 'Right Place, Wrong Person' Compilation G...	rms right place wrong person compilation gigat...
1	Weekly /r/bangtan Room (방탄방) - June 01, 2024 #...		
2	Jean, creator of @BTSMerchUpdates on X, has pa...		
3	240603 [Notice] 2024 FESTA BTS Jin Offline Eve...	240603 [Notice] 2024 FESTA BTS Jin Offline Eve...	240603 notice 2024 festa bts jin offline event...
4	240603 wootteo on Instagram [] alltherach_: tr...	240603 Wootteo on Instagram [] Alltherach_: Tr...	240603 wootteo instagram alltherach trans orr3...

Figure1. Cleaned Dataset

3.2.2. Data Cleaning:

The following steps were taken to clean the data:

Removing Emojis:

We removed all emojis instead of translating their meaning, as the context is already captured in the conversations, and emojis were frequent enough to interfere with analyzing top words in the text.

Replacing URLs with the Word "link":

We removed all URLs and replaced them with the word "link," allowing us to understand that a link was present after its removal.

Removing Hashtags and Special Characters:

We removed hashtags and any special characters because the data, collected directly from Twitter, was noisy.

Translating Text in Other Languages:

We translated any non-English text that appeared within the English text.

Removing Fandom Group Names:

We removed most fandom group and idol names as they interfered with analyzing top words. A manually curated list of names related to SF9 and others was used for this purpose.

Lowercasing:

All text was converted to lowercase for better analysis.

Tokenization:

The text was tokenized into words using the tokenizer from the NLTK library.

Removing Stop Words:

English stop words were removed using the NLTK library.

Lemmatization:

This process converts words to their root forms and compares them with dictionary words to ensure grammatical correctness for analysis.

Keyword Filtering:

We used a keyword list from last year's work, which included words related to social support. We focused mainly on emotional support, so most words in this list are related to emotional support, like "empathy," "love," and "like."

POS Tagging:

Part-of-speech tagging was performed to remove words like prepositions, conjunctions, determiners, pronouns, numbers, etc.

3.3. Text Representation

3.3.1 TF-IDF:

We calculated Term Frequency-Inverse Document Frequency to weigh the importance of words. This representation was used for all four topic models: Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), and BERTopic.

3.3.2. Voyage-AI:

Instruction-tuned general-purpose embedding model optimized for clustering, classification, and retrieval.

3.4. Topic Modeling

We applied various topic modeling techniques, including LDA, NMF, LSA, and BERTopic.

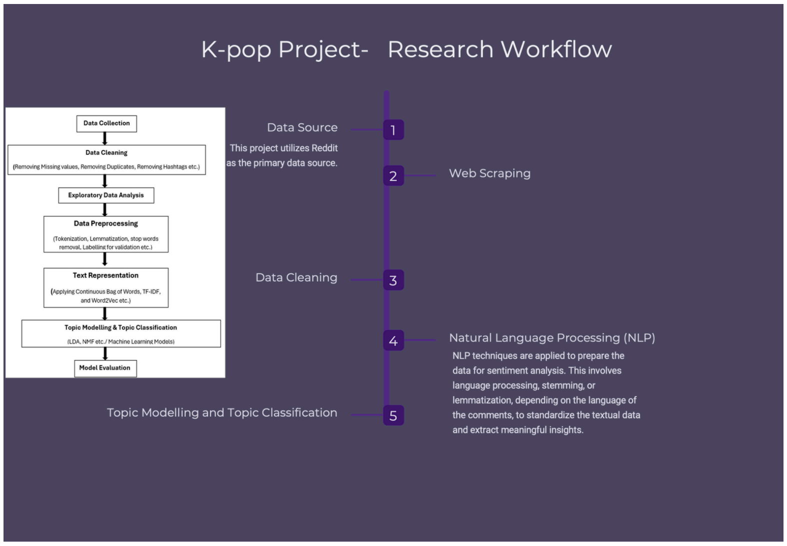


Figure1: Research Methodology

3.4.1. Latent Dirichlet Allocation (LDA):

A generative statistical model used for topic modeling that discovers latent topics in a set of documents by assuming each document is a mixture of topics and each topic is a mixture of words. LDA is suitable for this case as it performs random sampling of the data, providing better text analysis. [2]

3.4.2 Non-Negative Matrix Factorization (NMF):

A matrix factorization method that decomposes a document-term matrix into two lower-dimensional non-negative matrices to identify latent topics. Each document and topic are represented by non-negative combinations of topics and words, respectively. NMF is effective as it handles sparsity well. [3]

3.4.3 Latent Semantic Analysis (LSA):

A dimensionality reduction technique that decomposes the document-term matrix using Singular Value Decomposition (SVD) to discover latent relationships between terms and documents, often used for topic modeling. LSA is chosen for its ability to provide better semantic relationships among the text. [4]

3.4.4 BERTopic:

It is a topic modeling technique that utilizes Huggingface transformers and class-based TF-IDF to create dense clusters for easy interpretation of topics while keeping the important words in the topic description

BERTopic Architectures:

BERTopic with Voyage-AI Embeddings

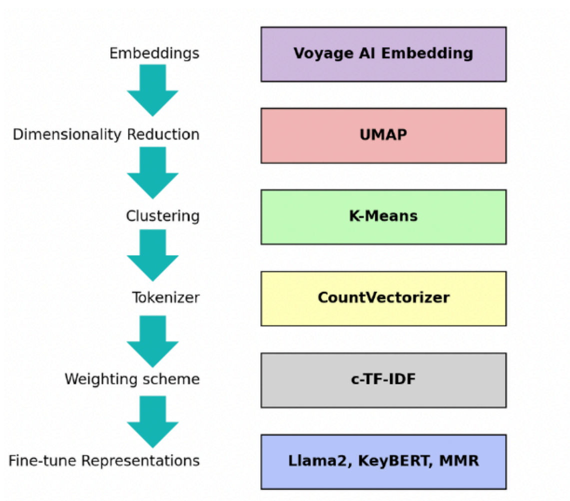


Figure 2a

BERTopic with TF-IDF Embeddings

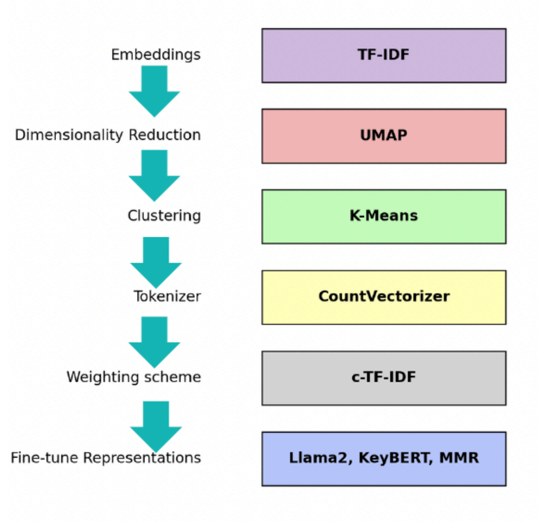


Figure 2b

3.4.4.1 Embeddings:

We start by embedding the documents i.e., the text extracted from twitter and reddit. The documents are converted into numerical representations. The default method for embeddings is sentence-transformers, which is optimized for semantic similarity that is important for clustering tasks. For our project we have used two types of embeddings i.e., TF-IDF and Voyage-AI embedding.

3.4.4.2 Dimensionality Reduction

After we have obtained the numerical representations of the document, the next task is to reduce the dimensions of these embeddings as the clustering models cannot handle the high dimensional data due to the curse of the dimensionality. The UMAP is the dimensionality reduction method used to reduce the dimensions of the embeddings. The reasons for using UMAP are:

- It is fast. It can handle large datasets and high dimensional data without too much difficulty, scaling beyond what most t-SNE packages can manage.
- UMAP scales well in embedding dimensions.
- UMAP often performs better at preserving some aspects of the global structure of the data than most implementations of t-SNE.
- UMAP supports a wide variety of distance functions, including non-metric distance functions such as cosine distance and correlation distance. We can embed word vectors properly using cosine distance.

3.4.4.3 Clustering

After we obtain the embeddings, we use clustering algorithms to get the clusters of the low dimensional embeddings. We used the k-means algorithm for the clustering.

K-means Clustering:

K-means clustering is a widely used technique for grouping data points by minimizing the mean squared distance to the nearest center. Lloyd's algorithm is a common heuristic for a simple and efficient implementation called the filtering algorithm. Unlike other methods, it precomputes a kd-tree for the data points. The algorithm's practical efficiency is demonstrated through a data-sensitive analysis and empirical studies on both synthetic and real data in applications like color quantization and image segmentation. [6]

3.4.4.4 Tokenizer (Bag of Words)

We combine all documents within a cluster into a single, long document to represent the cluster. By counting word frequencies in each cluster, we create a bag-of-words representation at the cluster level, rather than the document level. This approach focuses on word usage across topics (clusters) without assuming any specific structure. To account for varying cluster sizes, the bag-of-words representation is L1-normalized. For the tokenization task, we have used Countvectorizer.

3.4.4.5 Weighting Scheme (Topic Representation)

To distinguish between clusters, we can modify TF-IDF to focus on clusters instead of individual documents. By treating all documents in a cluster as a single document and applying TF-IDF, we obtain importance scores for words within each cluster. This approach highlights words that are most representative of a cluster's topic, effectively describing the topics through the most important words. This method is known as class-based TF-IDF:

c-TF-IDF

For a term **x** within class **c**:

$$W_{x,c} = \|tf_{x,c}\| \times \log\left(1 + \frac{A}{f_x}\right)$$

$tf_{x,c}$ = frequency of word **x** in class **c**

f_x = frequency of word **x** across all classes

A = average number of words per class

Figure 3. Mathematical formula of c-TF-IDF[5]

3.4.4.6 Fine-Tune Representations (Representation Models)

After generating c-TF-IDF representations, which provide a quick and accurate summary of topics, these initial topic representations can be further refined using advanced NLP methods like GPT, T5, KeyBERT, and SpaCy. In BERTopic, these c-TF-IDF generated topics serve as candidate topics, each with a set of keywords and representative documents. This setup allows for efficient fine-tuning, as large models only need to process a smaller set of documents per topic, making the use of models like GPT and T5 feasible in production and reducing overall computation time. For our representation model, we used Llama2 with 7 billion parameters (meta-llama/Llama-2-7b-chat-hf) and KeyBERT. [7]

To load the llama2 model, the transformer and PyTorch were used to load the model. Bits and Bytes configuration was used to load the model in 4 bit configuration. The original 7 billion parameter model required a high computational power which was countered by using a bits and bytes configuration. [5]

3.5. Evaluation

- **Coherence Score:** Measure the semantic coherence of the topics. High coherence scores indicate more interpretable topics.
- **Human Judgment:** Manually evaluate the topics for interpretability and relevance.
- **Topic Diversity:** Assess the diversity of the topics to ensure that the model captures a wide range of themes.
- **Visualization:** Use tools like pyLDAvis to visualize the topics and their distributions, helping to qualitatively assess the model.
- Used some metrics like perplexity for LDA, singular value analysis for LSA and reconstruction error for NMF.

4. RESULTS

4.1 Coherence Score of Topic Models

We have obtained the coherence scores for the various topic models that were used for the project.

Cv Coherence Score:

In c_v coherence, each topic word is compared with all topics using a boolean sliding window to assess co-occurrence. A word vector of size N is created, where each cell contains the Normalized Pointwise Mutual Information (NPMI) between the word and others. These word vectors are aggregated into a topic vector, and the c_v score is the average cosine similarity between each word and its topic vector. [8]

Cv ranges between 0 and 1.

Topic Model	C_v Coherence Score
Latent Dirichlet Allocation (LDA)	0.63
Non-Negative Matrix Factorization (NMF)	0.54
Latent Semantic Analysis (LSA)	0.37
BERTopic with Voyage AI Embeddings	0.59
BERTopic with TF-IDF Embeddings	0.44

Table 1. Coherence Score Comparison

4.2 Comparison of BERTopic Models

We have compared the BERTopic Models that we have used using the two types of Coherence Scores i.e., Cv Coherence Score and U_mass Coherence Score.

U_mass Coherence Score:

UMass coherence score, which measures how often two words, wi and wj, appear together in a corpus. It is defined as:

$$C_{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

Figure 4. Mathematical representation of U Mass Coherence Score [9]

where D(wi,wj) represents the co-occurrence frequency, and D(wi)is the frequency of wi alone. This asymmetric measure is averaged across the top N words of a topic to calculate global coherence, with higher values indicating better coherence. [9]

Metrics	**BERTopic with Voyage AI Embeddings**	**BERTopic with TF-IDF Embeddings**
U_mass Coherence Score	-3.10	-5.49
C_v Coherence Score	0.59	0.44

Table 2. Comparison of BERTopics

4.3 Top Words Bar Chart

The following plots show the top words in each topic with the scores obtained for both BERTopic with Voyage AI Embeddings and TF-IDF Embeddings.

Figure 4a. Top Words with Voyage AI Embeddings

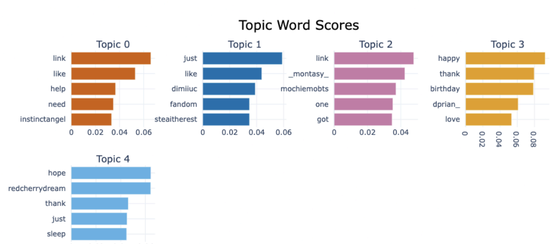
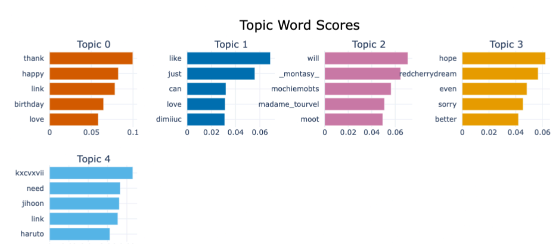


Figure 4b. Top Words with TF-IDF Embeddings



4.4 Word Cloud of BERTopic with Voyage AI Embeddings

The word cloud below represents the top words in each topic.

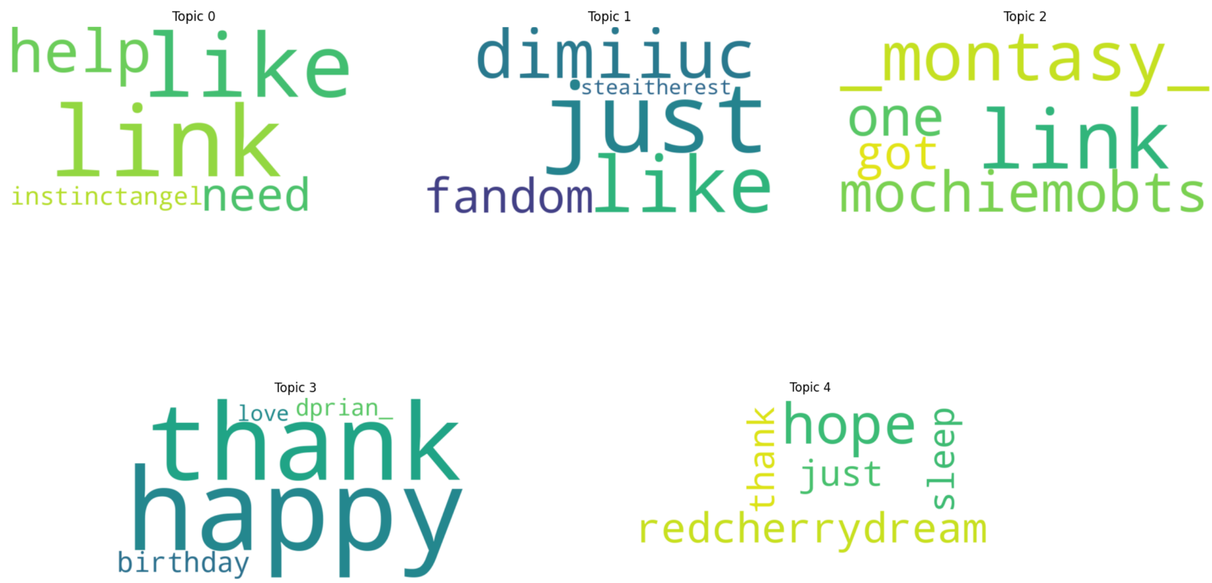


Figure 5. Word Cloud with Voyage AI Embedding

4.5 Word Cloud of BERTopic with TF-IDF Embeddings

The word cloud below represents the top words in each topic.

Figure 6a. Word Cloud with TF-IDF Embeddings

Figure 6b. Word Cloud with TF-IDF Embeddings

4.6 Comparison of Representation Models

Following tables contain the Representative Words for KeyBERT and Llama2 Representation for the Topics Obtained from BERTopic with Voyage-AI Embeddings and BERTopic with TF-IDF Embeddings.

Voyage-AI

	Llama2	KeyBERT				
Topic 0:	'Mental Health and the Use of Packing Tape'	'stuff'	'lf'	'smt'	'mean'	'like'
Topic 1:	'Fandom Discussions and Personal Experiences'	'fandom'	'feel'	'life'	'whatever'	'keep'
Topic 2:	"K-pop and K-beauty enthusiasts' online interactions and preferences"	'yooniquelyja'	'krklemm'	'kebinwooo'	'__ysyh'	'원어스'
Topic 3:	'Happy Birthday'	'birthday'	'day'	'happy'	'congratulation'	'life'
Topic 4:	'Mental Health and Sleep Deprivation'	'rest'	'break'	'sleep'	'feeling'	'hope'

TF-IDF

	Llama2	KeyBERT				
Topic 0:	'Happy Birthday'	'birthday'	'day'	'congratulation'	'happy'	'thankful'
Topic 1:	'Social Media Discourse and Online Fandom'	'fandom'	'tattoo'	'mean'	'feel'	'say'
Topic 2:	'K-pop fandom and merchandise'	'yooniquelyjae'	'tweet'	'soobinurhome'	'yootae'	'spcygomtokki'
Topic 3:	'Mental Health and Self-Care'	'rest'	'break'	'work'	'sleep'	'feeling'
Topic 4:	'K-pop and social media culture'	'kxcvxviii'	'handsome'	'tomorrowbyskz'	'yall'	'jihoon'

Figure 7. Representation Models

5. DISCUSSION

5.1 Discussion of results

In the project as outlined in the methodology, we have shown a comparison of various topic models for the sentiment analysis and finding evidence of social support in the data extracted from the twitter which was for sf9 fandom (K-pop fandom group). The study started with defining the problem statement for the project, collecting the data, doing literature survey, brainstorming various approaches for the topic modeling, defining and implementing the approach and evaluating the results.

The topic models LDA, NMF, LSA, and BERTopic performed decently given the size of the dataset. LDA performed the best with a coherence score of 0.65, and LSA being the worst with a coherence score of 0.37 as highlighted in section 4.1. With BERTopic, the coherence score obtained was 0.59 which is a decent score considering the small size of the dataset with only 158 documents.

In the discussion our primary focus would be on the BERTopic. We are showing comparative study of BERTopic with Voyage-AI Embeddings and TF-IDF Embeddings. In section 4.2, we have compared both the BERTopic models using Cv and U_mass Coherence scores. The table shows that Voyage-AI has performed better in both the metrics with U-mass score of -3.1 and Cv score of 0.57. The reason for Voyage-AI performing better is that it has captured the semantic meaning between the documents better than the TF-IDF embeddings.

Topic Modeling Results

- **Topic Coherence**:
  - **Voyage AI**: The topics generated using Voyage AI embeddings appear more coherent, with words within each topic being semantically related. This indicates that Voyage AI embeddings can effectively cluster words that share a common theme or sentiment.
  - **TF-IDF**: The topics generated using TF-IDF embeddings show a mix of relevant and less relevant words, which can dilute the overall coherence of the topic. For

instance, words like "can," "like," and "love" might not belong together in a single sentiment-focused topic.

- **Word Clouds**:
  - **Voyage AI Word Clouds**: The word clouds generated using Voyage AI embeddings present a more focused set of keywords, which can help in better understanding the main themes within each topic.
  - **TF-IDF Word Clouds**: The word clouds from TF-IDF embeddings might contain more noise, as high-frequency words that are not contextually significant may still dominate the representation.

5.2 Interpretability and Insights

- **Voyage AI**:
  - **Deeper Insights**: By leveraging more sophisticated embeddings, Voyage AI allows for deeper insights into the data, making it easier to draw conclusions about the underlying sentiments and themes.
  - **Use in Dynamic Text**: Voyage AI is better suited for dynamic, context-rich text like social media posts, where the meaning of words can change based on the context they are used in.
- **TF-IDF**:
  - **Broad Insights**: TF-IDF provides broader insights, which might be more suitable for tasks where a general overview is sufficient rather than a deep dive into the nuances of the text.

    - **Scalability**: TF-IDF is computationally less intensive and easier to scale, making it a good choice for larger datasets where fine-grained analysis may not be necessary.

Llama2 and KeyBERT Comparisons

- **Voyage AI with Llama2 and KeyBERT**:
    - Shows more varied and contextually relevant topics, which can be beneficial for identifying nuanced trends in the data.
    - The outputs reflect specific themes, such as "Mental Health and Sleep Deprivation," which are more targeted and useful for detailed analysis.
- **TF-IDF with Llama2 and KeyBERT**:
    - Presents broader themes that might miss some of the finer details captured by Voyage AI embeddings.
    - The output themes, while relevant, are less specific, which might limit their utility in applications requiring detailed sentiment analysis.

5.3 Conclusion

In summary, Voyage AI embeddings provide a more sophisticated and context-aware analysis, which is particularly useful for sentiment analysis and topic modeling in complex and dynamic text data. TF-IDF embeddings, while simpler and more general, might be better suited for large-scale, less detailed analyses. The choice between these methods should be guided by the specific needs of the analysis—whether the goal is to capture fine-grained nuances or to achieve a broader, more general understanding of the text data.

6. LITERATURE CITED

[1] Grootendorst, Maarten. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." *arXiv preprint arXiv:2203.05794*, 2022, https://arxiv.org/abs/2203.05794.

[2] Zhao, Siqi. "Latent Dirichlet Allocation (LDA)." *Towards Data Science*, Medium, 24 Jan. 2018, towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2. Accessed 12 Aug. 2024.

[3] Prasad, Jay. "What Is Non-Negative Matrix Factorization (NMF)?" *Codex*, Medium, 8 July 2020,medium.com/codex/what-is-non-negative-matrix-factorization-nmf-32663fb4d65. Accessed 12 Aug. 2024.

[4] "Latent Semantic Analysis." *Wikipedia*, The Free Encyclopedia, 12 Aug. 2024, en.wikipedia.org/wiki/Latent_semantic_analysis. Accessed 12 Aug. 2024.

[5] Maarten Grootendorst. "BERTopic: Leveraging BERT for Topic Modeling." *Maarten Grootendorst*, maartengr.github.io/BERTopic/index.html. Accessed 12 Aug. 2024.

[6] Kanungo, Tapas, et al. "The Analysis of a Simple k-Means Clustering Algorithm." *Computational Geometry 2000 Hong Kong China*, ACM, 2000, doi:10.1145/336154.336189. Accessed 12 Aug. 2024.

[7] Meta. "Llama 2 7B Chat HF." *Hugging Face*, huggingface.co/meta-llama/Llama-2-7b-chat-hf. Accessed 12 Aug. 2024.

[8] O'Callaghan, Joe. "c□: Topic Coherence Explained." *Towards Data Science*, Medium, 26 Oct. 2021, towardsdatascience.com/c□-topic-coherence-explained-fc70e2a85227. Accessed 12 Aug. 2024.

[9] Gullberg, Peter. "Topic Modeling Coherence Score." *Baeldung*, 2 Feb. 2023, www.baeldung.com/cs/topic-modeling-coherence-score. Accessed 12 Aug. 2024.

7. ADDITIONAL FILES

https://github.com/sgunpreetsandhu01/The-Globalizing-K-pop-Project-Sentiment-Analysis-of-Global-Fandoms-on-Social-Media-.git