INTERNSHIP REPORT

SUBMITTED BY-

ADITI SATSANGI
Globalink Mitacs Research Intern

UNDER THE SUPERVISION OF-

Dr. Stacey Scott
Dr. Ritu Chaturvedi

Mitacs Globalink Research Internship 2024

School of Computer Science

University of Guelph

The Globalizing K-Pop Project- Analysing Social Support in K-Pop Fandoms on Social Media Using Topic Modelling and Large Language Models

Contents

Acknowledgement

I would like to express my sincere gratitude to the following people for their help and support in the completion of this project:

- Our project supervisor, **Dr. Stacey Scott**, Professor at University of Guelph for her guidance and feedback throughout the internship.
- Our project supervisor professor **Dr. Ritu Chaturvedi** for her guidance and support throughout the internship.
- Our Team, Gunpreet and Krish Garg, for their help with brainstorming and research.
- The **[library/database/website]** for providing me with access to valuable resources.

I would also like to thank the School of Computer Science, University of Guelph for providing us with the opportunity to work on this project.

The Globalizing K-Pop Project- Analysing Social Support in K-Pop Fandoms on Social Media Using Topic Modelling and Large Language Models

Abstract:

This project is a comprehensive study aimed at understanding the role of social media in enhancing well-being and fostering a sense of community among global K-pop fandoms. Through sentiment analysis using topic modelling techniques, this research examines data extracted from Twitter and Reddit, focusing on various forms of social support, including emotional, informational, and appraisal support 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 analyse the collected data, which has been pre-processed through tokenization, lemmatization, and keyword filtering.

The research workflow involves extracting top keywords, visualizing clusters, and utilizing representation models such as Llama2, BART, GPT, and KeyBERT to present the results to 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 a commendable Cv coherence score of 0.59, despite the limited dataset available. Among the models tested, Non-Negative Matrix Factorization (NMF) demonstrated the highest effectiveness, yielding a Cv coherence score of 0.656. The Latent Dirichlet Allocation (LDA) model produced a moderate score of 0.59, while the Latent Semantic Analysis (LSA) model exhibited the lowest coherence score of 0.416. These results underscore the relative strengths and limitations of each modelling approach within the context of the data used.

Introduction:

Sentiment analysis, an essential tool for 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 modelling, a method traditionally used to uncover 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]

Applying topic modelling techniques to understand social support in the K-pop fandom community offers promising insights. Large Language Models (LLMs) are AI models trained on vast amounts of text data to understand and generate human language. These models are employed in this study to represent the results of topic models, providing a more comprehensive analysis of the sentiments and themes present within K-pop fandoms.

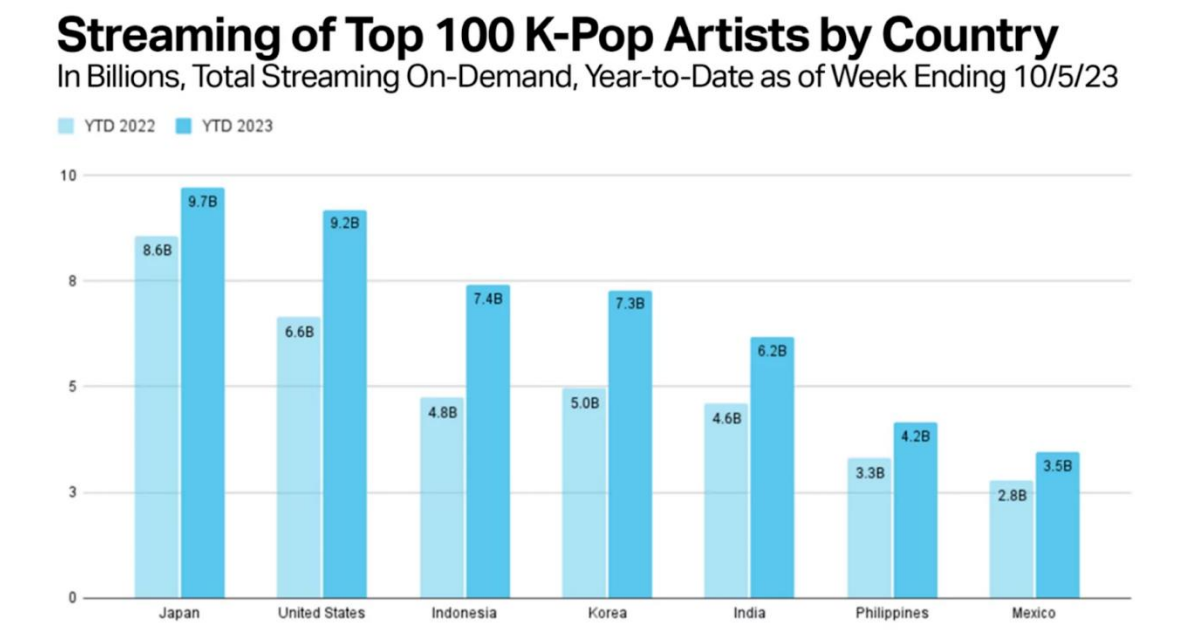


Fig 1. Depicting the increasing trend of the K-Pop in different countries.

Methodology:

➢ Data Collection

1. **Reddit:** Collect comments from Reddit focusing on reviews related to K-pop.

   Targeted K-Pop Music:

   - Bangtan Boys (BTS)
   - Blackpink
   - SF9
   - Twice

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

2. **Twitter Data:** We got twitter data which was collected using twitter API in 2023. This data is related to Sf9 fandom. Another twitter data on which we worked was related to the K-Pop which we found on GitHub. The GitHub link is shown below:

   https://github.com/tsainez/kpop-sentiment-analysis/blob/main/data/data.csv

The preview of twitter dataset (from GitHub) is given below:

```
df.head()
```

	id	text	created_at	lang
0	1529697716368211968	에버랜드 X HYBE 가든 오브 라이츠 투바투 개쩔어유 어머뽈자 영원럽 G...	2022-05-26 05:35:30+00:00	ko
1	1529697654334115842	Hybe successfully take the spotlight from Gar...	2022-05-26 05:35:15+00:00	en
2	1529697612659470336	jnklaina Im literally shock as well IDK if it...	2022-05-26 05:35:05+00:00	en
3	1529697533580427265	GIVEAWAY SPECIAL TH ANNIVERSARY SEVENTEEN CAR...	2022-05-26 05:34:47+00:00	en
4	1529697452336369664	Thank god hybe doesnt gaf about yalls ridicul...	2022-05-26 05:34:27+00:00	en

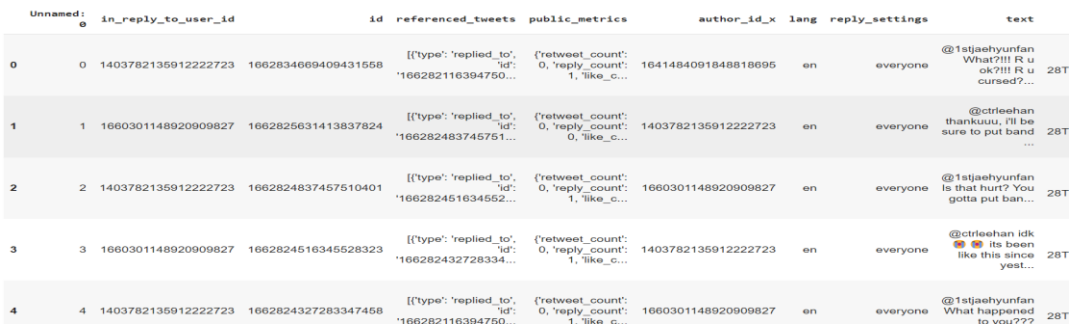
The link for recently used twitter sf9 data (un-preprocessed) used for final results is shown below:

https://gitlab.socs.uoguelph.ca/sscott15/social-support-in-kpop-fandoms/-/blob/Project/Dataset/SF9_fanclub_conv_followers_100_500.xlsx?ref_type=heads

The link for pre-processed sf9 data used for final results is given below:

https://gitlab.socs.uoguelph.ca/sscott15/social-support-in-kpop-fandoms/-/tree/Project/Dataset/Preprocessed%20Data-%20Recently%20used%20twitter%20sf9?ref_type=heads

The preview of sf9 twitter recent dataset is given below:



	Unnamed: 0	in_reply_to_user_id	id	referenced_tweets	public_metrics	author_id_x	lang	reply_settings	text	
0	0	1403782135912222723	1662834669409431558	[{'type': 'replied_to', 'id': '166282116394750...	{'retweet_count': 0, 'reply_count': 1, 'like_c...	1641484091848818695	en	everyone	@1stjaehyunfan What?!!! R u ok?!!! R u cursed?...	28T
1	1	1660301148920909827	1662825631413837824	[{'type': 'replied_to', 'id': '166282483745751...	{'retweet_count': 0, 'reply_count': 0, 'like_c...	1403782135912222723	en	everyone	@ctrleehan thankuuu, i'll be sure to put band ...	28T
2	2	1403782135912222723	1662824837457510401	[{'type': 'replied_to', 'id': '166282451634552...	{'retweet_count': 0, 'reply_count': 1, 'like_c...	1660301148920909827	en	everyone	@1stjaehyunfan Is that hurt? You gotta put ban...	28T
3	3	1660301148920909827	1662824516345528323	[{'type': 'replied_to', 'id': '166282432728334...	{'retweet_count': 0, 'reply_count': 1, 'like_c...	1403782135912222723	en	everyone	@ctrleehan idk 😭😭 its been like this since yest...	28T
4	4	1403782135912222723	1662824327283347458	[{'type': 'replied_to', 'id': '166282116394750...	{'retweet_count': 0, 'reply_count': 1, 'like_c...	1660301148920909827	en	everyone	@1stjaehyunfan What happened to you???	28T

Fig 2. Preview of the dataset collected from Twitter (Sf9)

➢ Data Preprocessing

1. Linking Conversation:

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".

```
conversation_df.head()
```

	conversation_id	text	reply_count	first_tweet
0	1658489916412801024	Fantasy friends plz help me to find the nicest...	2	Fantasy friends plz help me to find the nicest...
1	1658831414291226624	praying that kflex london has an amazing lineu...	1	praying that kflex london has an amazing lineu...
2	1658872268708265984	im already tired but still need to stay up to ...	1	im already tired but still need to stay up to ...
3	1658916239949930553	I will probably delete this tweet later but I ...	1	I will probably delete this tweet later but I ...
4	1659090060279529472	everyone pray for my accounting tmrw...1139746...	5	everyone pray for my accounting tmrw

Fig 3. Preview of Dataset after linking the conversations

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 analysing top words in the text.
   - **Replacing the URLs with word 'link':**
     - We removed all URLs or the links and replaced them with word "link". As this word will help us to understand that the link was there after removal of the link.
   - **Removal of Hashtags and the Special Characters:**
     - We removed the hashtags and the any kind of the special characters as we collected the data from the twitter directly, so the data was completely full of the noise.
   - **Translation of other languages text:** We translated the text of other languages if it occurs in the English text also.

   - **Removing the Fandom Group names**: We removed the mostly fandom groups and idols names as they were also creating the problem while analysing the top words in the results. Manually I prepared the list of the names of these groups and idols related to sf9 and others also.
   - **Lowering case**: We lowered the cases of text for better analysis.
   - **Tokenization:** We converted the text into words using the tokenizer from the NLTK library.

   - **Removal of Stop Words:** We removed the English stop words form the text by using 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.

After all these procedures, we obtained the pre-processed dataset whose preview is given below. The column “Text” and “cleaned_text” are almost same, just difference lies in the words and sentence form. The “Text” column includes the text in terms of the words. The “cleaned_text” column include the same text in the form of sentence.

```
filtered_df.head()
```

	conversation_id	text	reply_count	first_tweet	Text	cleaned_text
0	1658489916412801024	Fantasy friends plz help me to find the nicest...	2	Fantasy friends plz help me to find the nicest...	[fantasy, friend, plz, help, find, nicest, ful...	fantasy friend plz help find nicest full body ...
1	1658831414291226624	praying that kflex london has an amazing lineu...	1	praying that kflex london has an amazing lineu...	[praying, kflex, london, amazing, lineup, just...	praying kflex london amazing lineup just like ...
2	1658872268708265984	im already tired but still need to stay up to ...	1	im already tired but still need to stay up to ...	[already, tired, still, need, stay, assignment...	already tired still need stay assignment help ...
3	1658916239949930553	I will probably delete this tweet later but I ...	1	I will probably delete this tweet later but I ...	[will, probably, delete, tweet, later, just, g...	will probably delete tweet later just got inte...
4	1659090060279529472	everyone pray for my accounting tmrw...1139746...	5	everyone pray for my accounting tmrw	[everyone, pray, accounting, tmrw, aww, thank,...	everyone pray accounting tmrw aww thank bff ri...

Fig 4. Preview of the Pre-Processed Dataset

➢ Text Representation

1. **TF-IDF:** Calculated the Term Frequency-Inverse Document Frequency to weigh the importance of words. We used this for our all four topic models Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA) and BERTopic.
2. **Voyage-AI:** Instruction-tuned general-purpose embedding model optimized for clustering, classification, and retrieval

➢ Topic Modelling

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

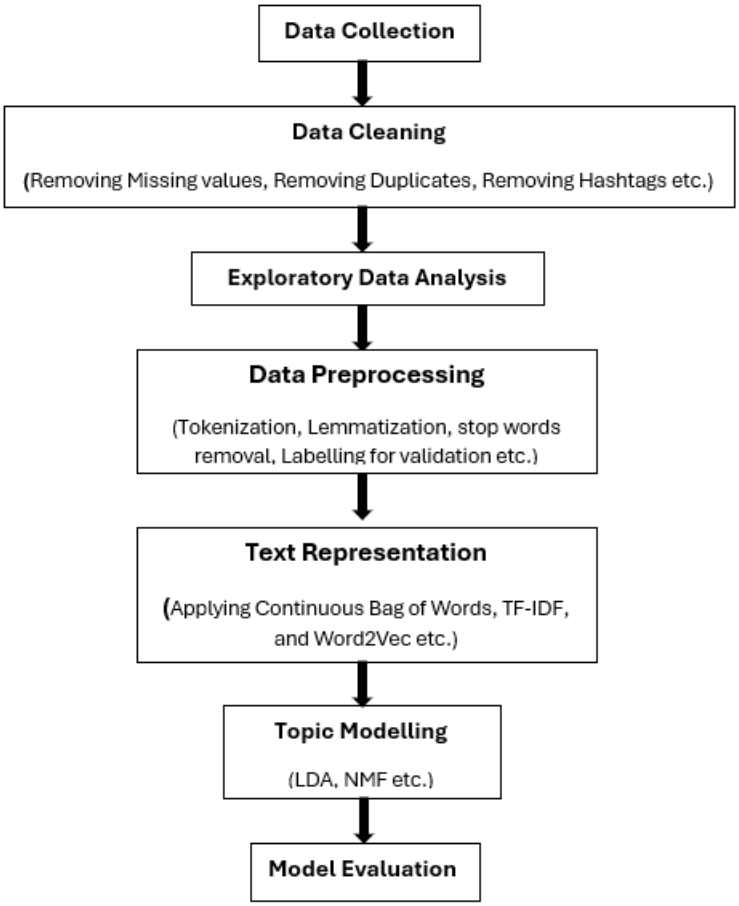


Fig 5: Research Methodology

1. **Latent Dirichlet Allocation:**
   A Latent Dirichlet Allocation (LDA) model is a generative statistical model used for topic modelling 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 used for this case as it performs the random sampling of the data which provides the better analysis of the text [2].

2. **Non-Negative Matrix Factorization:**
   A matrix factorization method that decomposes a document-term matrix into two lower-dimensional non-negative matrices to identify latent topics where each document and topic are represented by non-negative combinations of topics and words, respectively. We used NMF as

it works well with sparsity as well [3].

The objective function is:

$$
\begin{aligned}
L(W,H) = {} & 0.5 * ||X - WH||^2_{loss} \\
& + alpha\_W * l1\_ratio * n\_features * ||vec(W)||_1 \\
& + alpha\_H * l1\_ratio * n\_samples * ||vec(H)||_1 \\
& + 0.5 * alpha\_W * (1 - l1\_ratio) * n\_features * ||W||^2_{Fro} \\
& + 0.5 * alpha\_H * (1 - l1\_ratio) * n\_samples * ||H||^2_{Fro}
\end{aligned}
$$

Where:

$||A||^2_{Fro} = \sum_{i,j} A^2_{ij}$ (Frobenius norm)

$||vec(A)||_1 = \sum_{i,i} abs(A_{ij})$ (Elementwise L1 norm)

The generic norm ||X−WH|| loss may represent the Frobenius norm or another supported beta-divergence loss. The choice between options is controlled by the beta_loss parameter. The regularization terms are scaled by n_features for W and by n_samples for H to keep their impact balanced with respect to one another and to the data fit term as independent as possible of the size n_samples of the training set. The objective function is minimized with an alternating minimization of W and H.

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 modelling. We used LSA as it provides the better semantic relationship among the text [4].

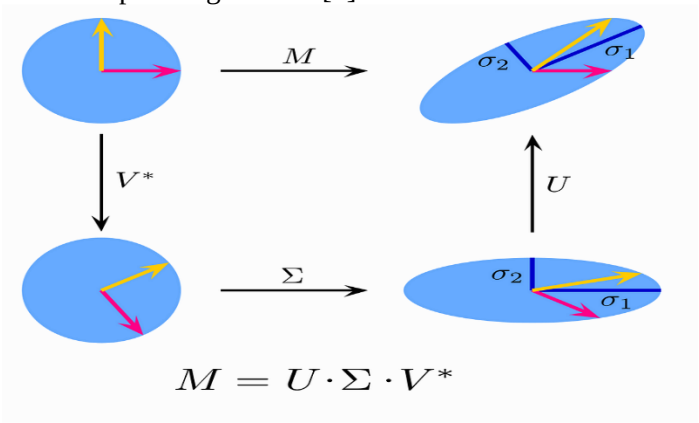


Fig. 6; Singular Value Decomposition

where *M* is our original (*m, n*) data matrix — m rows, n columns, *m documents, n terms*

U is a (*m, r*) matrix — *m documents and r concepts*

Σ is a *diagonal* (*r, r*) matrix — all values except those in the diagonal are zero. (But what do the non-zero values represent?

V is a (*n, r*) matrix — *n terms, r concepts*

The values in **Σ** represent how much each latent concept explains the variance in our data. If we were to decompose this to 5 components, this would look something like this:

$$M_{trunc} = \sum_{i}^{5} \sigma_i \underline{u}_i \otimes \underline{v}_i^T$$

A sum of the outer product of our weighted document-concept vector and our term-concept vector

where there would be originally *r* number of *u* vectors; 5 singular values and n number of *v-transpose* vectors.

4. **BERTopic:**

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

- **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 10 tasks. For our project we have used two types of embeddings i.e., TF-IDF and Voyage-AI embedding.
- **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.
- **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.
- **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.
- **Weighting Scheme** (Topic Representation) To distinguish between clusters, we can modify TF-IDF to focus on clusters instead of individual documents. [8]

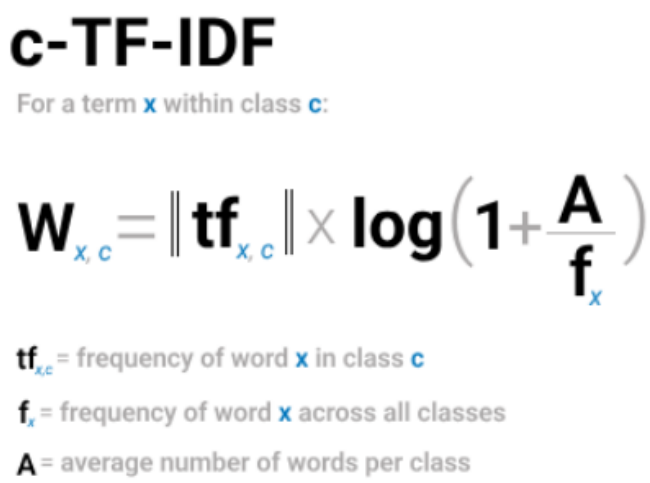


Fig. 7. Mathematical Formula for c-TFIDF

- **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.

➢ 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 ans reconstruction error for NMF.

➢ Results

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 14 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.

Table 1. Comparison of Coherence Scores of different models

Topic Model	Coherence Score
LDA	0.599
NMF	0.656
LSA	0.416
BERTopic with Voyage AI	0.59
BERTopic with TF-IDF	0.44

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:

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

Fig 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

3. Other Evaluation Metrics for LDA, NMF, LSA:

We calculated the perplexity for the LDA model, reconstruction error for NMF and the singular values for LSA model for evaluation.

- **Perplexity**:
  Perplexity is a commonly used metric to evaluate the performance of topic models, including LDA. It measures how well the model predicts unseen or held-out documents. A lower perplexity score indicates better model performance.
  Lower perplexity scores indicate that the model can better predict the words in unseen documents, suggesting a better understanding of the underlying topics. However, it's essential to note that perplexity is not the only measure of topic model quality, and it should be considered alongside other evaluation metrics, such as coherence and human interpretation of topics.[5]

- **Reconstruction Error:**
  The reconstruction error (RecError) and relative error (RelChange, the amount of change from the reconstruction error in the previous step) can be used to diagnose whether the calculation is converging or not. [6]

- **Singular values:**

Singular values represent the importance of each dimension in the new, reduced space. They are essentially weights assigned to the new features.
Given a matrix A, its SVD is a factorization of the form:

A = U Σ V^T

Where:
U is an m x m orthogonal matrix (columns are orthonormal eigenvectors of AA^T)
Σ is an m x n diagonal matrix with non-negative real numbers on the diagonal (these are the singular values)
V^T is the transpose of an n x n orthogonal matrix (columns are orthonormal eigenvectors of A^TA).[7].

We apply Singular Value Decomposition to decompose the term-document matrix into three matrices: *U* (term space), Σ (diagonal matrix of singular values), and *V to the powerT* (document space).

Table 3: Other Evaluation Results from different models

Model	Metrics
LDA	Perplexity: -13.14
NMF	Reconstruction Error: 11.3%
LSA	Singular values for each topic: 2.94, 1.79, 1.74,1.62,1.52

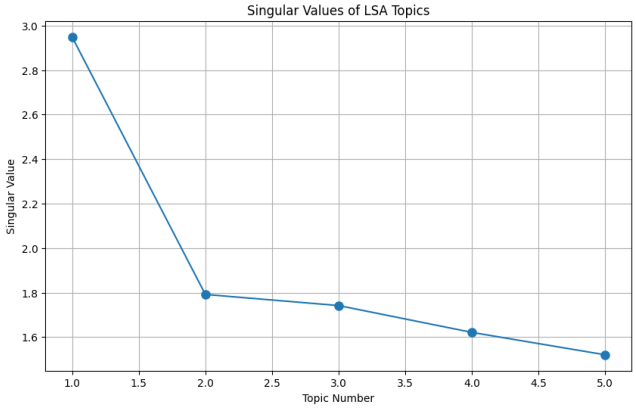


Fig 8. Singular values for different topics

4. 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

5. pyLDAvis Visualization:

PyLDAvis is a Python library used to visualize LDA (Latent Dirichlet Allocation) topics. It provides an interactive interface to explore the most probable words for each topic, as well as how documents are distributed across topics. This visualization is crucial for understanding the underlying semantic structure of a corpus and assessing the quality of the LDA model. It can also be generated for NMF after indirectly providing the component to the visualization tool. The fig. 9 represents the visualization by NMF model for topic 3 which is related to the requests for help. Deployed link for the visualization is given below:

https://aditisatsangi.github.io/Globalizing-K-Pop-Project-Analysing-Social-Support-using-Topic-Modelling-and-LLMs/#topic=0&lambda=1&term=

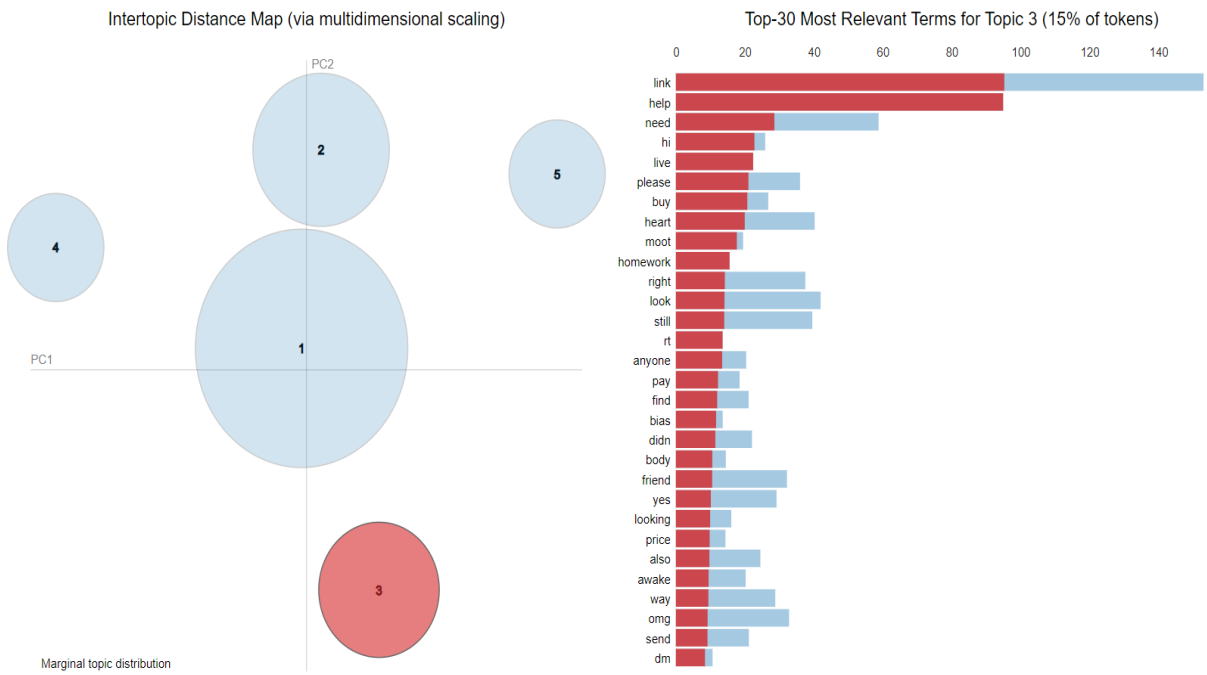


Fig. 9 pyLDAVis Visualization for NMF model

The fig. 10 represents the visualization by LDA model for topic 3 which is related to the fan content and interactions.

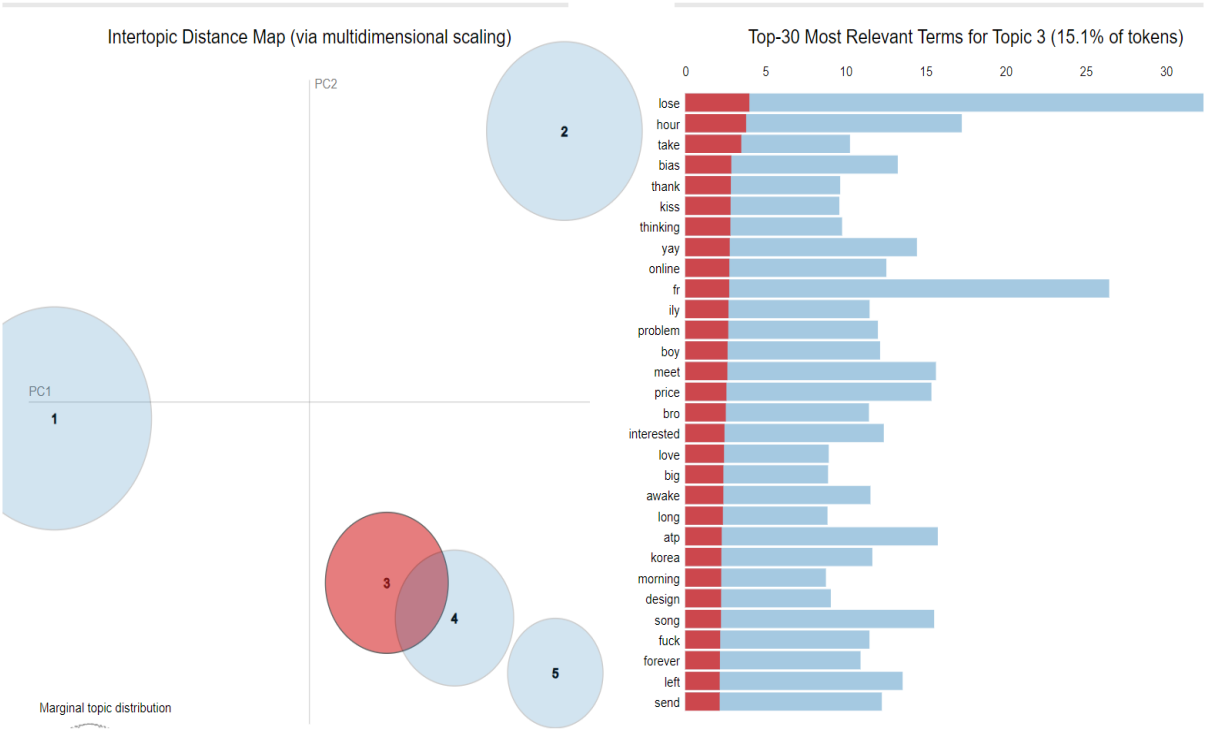


Fig. 10 pyLDAVis Visualization for LDA model

6. **<u>Top Words from different models:</u>** Top words given by different models are given below:

Top Wors from NMF:

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11
Topic 1	just	better	time	re	even	think	now	hope	sleep	feel	need
Topic 2	thank	happy	birthday	love	link	much	hope	cute	congrats	thanks	congratulation
Topic 3	link	help	need	hi	live	please	buy	heart	moot	homework	right
Topic 4	good	morning	song	feeling	link	great	today	ily	fun	luck	look
Topic 5	miss	back	emotional	tweet	fantasy	much	today	twitter	show	real	true

Top Words from LSA:

```
Topic 0:
link thank just love good time help hope much re

Topic 1:
thank birthday happy good love cute morning luck next concert

Topic 2:
link help morning buy heart live good hi need moot

Topic 3:
good morning song feeling look sleep great luck ily fun

Topic 4:
miss love morning heart want real pretty emotional fandom song
```

Top Words from LDA:

```
Topic 0: 0.006*"lose" + 0.005*"hour" + 0.005*"take" + 0.004*"bias" + 0.004*"thank" + 0.004*"kiss" + 0.004*"thinking" + 0.004*"yay" + 0.004*"online" + 0.004*"fr"

Topic 1: 0.007*"fr" + 0.006*"lose" + 0.005*"smile" + 0.005*"face" + 0.005*"cry" + 0.005*"price" + 0.005*"money" + 0.004*"tbh" + 0.004*"brazilian" + 0.004*"finished"

Topic 2: 0.011*"lose" + 0.008*"fr" + 0.005*"tbh" + 0.005*"bc" + 0.005*"cuz" + 0.004*"thankful" + 0.004*"gon" + 0.004*"lin" + 0.004*"atp" + 0.004*"break"

Topic 3: 0.007*"lose" + 0.005*"put" + 0.005*"face" + 0.005*"meet" + 0.004*"dont" + 0.004*"isn" + 0.004*"sleep" + 0.004*"kpop" + 0.004*"bias" + 0.004*"want"

Topic 4: 0.007*"break" + 0.006*"talking" + 0.005*"cuz" + 0.005*"honestly" + 0.005*"face" + 0.005*"lose" + 0.004*"today" + 0.004*"someone" + 0.004*"look" + 0.004*"met'
```

Fig. 11. Top Words of different topic models

7. **<u>Represeantation using Large Language Models (LLMs):</u>**

To enhance result interpretation, we leveraged several LLMs. KeyBERT was employed for keyword extraction from each topic, while BART generated concise topic summaries. GPT and Llama 2 were utilized for topic labelling, with the latter specifically integrated into the BERTopic model. We used the KeyBERT, GPT and ART for the representation of the results of the LDA, NMF and LSA model.

Table 4: Results from Representation Models for NMF Model

Topic	KeyBERT	GPT	BART
1	sleep, feel, hope, time, think	Emotional Support and Personal Struggles	look coworker face found...
2	birthday, congrats, cute, happy, thanks	Celebrations and Achievements	got birthday balloon believe...
3	heart, need, homework, moot, buy	Requests for Help	help link gunwookeg help...
4	morning, today, song, link, luck	Greetings and Daily Updates	good morning people...
5	twitter, tweet, fantasy, real, miss	Missing and Nostalgia	hope yall miss dumb tweet...

Table 5: Results from Representation Models for LSA Model

Topic	KeyBERT	GPT	BART
1	time, love, thank, help, link	Emotional Support and Comfort	look coworker face found...
2	concert, birthday, morning, happy, thank	Celebratory Moments and Birthdays	got birthday balloon believe...
3	moot, heart, morning, need, buy	Help and Requests	help link gunwookeg help...
4	morning, sleep, song, luck, feeling	Morning Greetings and Positivity	good morning people...
5	song, fandom, heart, emotional, miss	Nostalgia and Farewell	hope yall miss dumb tweet...

Table 6: Results from Representation Models for LDA Model

Topic	KeyBERT	GPT	BART
1	kiss, online, fr, hour, bias	Casual and Emotional Support	better best bl time...
2	brazilian, fr, lose, smile, price	Event Information and Seeking Help	still looking tix colosseum...
3	atp, break, fr, lose, bc	Fan Content and Interaction	rating vape flavour bc...
4	kpop, bias, sleep, face, lose	Personal Updates and Experiences	new addition family hehe...
5	break, met, talking, face, lose	Daily Life and Random Thoughts	idea sleep tje toilet...

Voyage AI:

	Llama2
Topic 0:	'Mental Health and the Use of Packing Tape'
Topic 1:	'Fandom Discussions and Personal Experiences'
Topic 2:	"K-pop and K-beauty enthusiasts' online interactions and preferences"
Topic 3:	'Happy Birthday'
Topic 4:	'Mental Health and Sleep Deprivation'

TF-IDF:

	Llama2
Topic 0:	'Happy Birthday'
Topic 1:	'Social Media Discourse and Online Fandom'
Topic 2:	'K-pop fandom and merchandise'
Topic 3:	'Mental Health and Self-Care'
Topic 4:	'K-pop and social media culture'

Fig 12. Results from the representation model for BERTopic

8. Top Documents from top performing model NMF:

Some top documents form each topic are given below:

Topic 1: Emotional Support and Personal Struggles:
Post: it's been a really rough day for me mentally so i just want to thank all of you who gave me birthday wishes today. i think i would go crazy without you guys. all i can do is hope tomorrow is better. i love you guys....

Responses:

Response 1: -@jaengpup i love u and i hope things get better for you soon https://t.co/LQMzBNUr6b...

Response 2: - @jaengpup no problem <3 https://t.co/mkWOjkTtPd...

Response 3: - @ayseetemasu my ayse thank u for always being kind to me...

Response 4: - @bearyuns i love u too bee :(...

Response 5: - @oxshyuk thank you so much eli :(this is really sweet...

Response 6: - @jaengpup we love you pup! tomorrow is going to be better and we will ALL always be behind you 100% ! be kind to yourself you deserve it...

Response 7: - @jaengpup we love you too pup 💟...

Response 8: - @jaengpup love u more than u know my jaspie 💟 https://t.co/Hz8ctY1bOh

Topic 2: Celebrations and Achievements:

Post: down and disheartened by things today, but i made it to 24. (i'm also spending it at work rip) i'm still on my break, but i wanted to check in. hope everyone is well. i'll be back with a new fic this month as a gift to you all. i hope u enjoy it. <3 be back soon (´ ω `)ノ https://t.co/ba6JNDP9Up

Responses:

Response 1 - @sonderstarlight Happy Happy … enjoy https://t.co/fTtkyznHVZ...

Response 2- @sonderstarlight Happy birthday Siren!! 💟 💟 🎉 Sending best wishes your way, may this year bring you many moments of love and happiness. Treat yourself, you deserve it- take care https://t.co/O4atCzDBfc...

Response 3- @sonderstarlight happy birthday!!...1328578923580502016- @sonderstarlight Happy happy day to you 🤍 🤍 🤍 🤍 🤍 🤍 🤍 🤍...1

Response 4- @sonderstarlight happy bday bb!!! hope you feel better soon and get lots of hugs and rest. stay hydrated!...

Response 5- @sonderstarlight HAPPY BALLOONY DAY SIREN!!!! 😎🥳🥳🥳🥳🥳💮💮💮💮💮💮💮💮💮 Please take some rest and enjoy your favourites if you can!...

Response 6- @sonderstarlight happy birthday! 🎉 I know birthdays can be rough sometimes, but I hope there are some moments you can enjoy!...

Response 7- @sonderstarlight happy birthday💕! i wish you lots of happiness and i hope you enjoy your day! 💟...

Topic 3: Requests for Help:

Post : HELP I AM NOT OKAY OMG https://t.co/n3KW0UTznp...

Responses:

Response 1 - @Lillinn333 Best way to start your day 😎...

Response 2- @kxcvxvii I woke up to so much hotness 😳😳 I'm melting...

Response 3- @kxcvxvii 😅😅😅😂😂😂😂😂 I think youngbin shocked a lot of his wife tonight....

Response 4- @keyz1206 He needs to make me come back to life with a kiss😭 😭 😭 😭...

Response 5- @kxcvxvii 😅😅😅😅😂😂😂😂 I will tell him in the fancafe that he gave you a heart attack 😂 😂 😂...

Response 6- @keyz1206 YES KEYZ HELP OUR HUSBAND IS KILLING ME https://t.co/cY4RKrVxM4...
Response 7- @kxcvxvii do you need an ambulance...
Response 8- @kxcvxvii 💧Hottie😳😜 https://t.co/yefxyy9uCL

Topic 3: Requests for Help:
Post: Still looking for 1 Tix Johnny Be Colosseum for my friend because her ticket got cancelled. She's a johfam so please help her to meet Johnny 🥺 wtb johnny be colosseum #JOHNNYbeAtColosseumJKT...

Responses:
Response 1 - @Midsummer_JY Hey why not message (Zaharaleonhard_) on Instagram she's still selling her ticket's...
Response 2 - @Midsummer_JY Hiii Message (_mary.roland_) on instagram, she's selling her tickets if you're still interested...
Response 3 - @Midsummer_JY MesAge me please can show proof of purchase and willing too sale for face value or less than face value...
Response 4- @Midsummer_JY Dm me, I'm looking to sell

Topic 4: Greetings and Daily Updates:
Post: good morning people in my phone ily https://t.co/n0hm2NtCc4

Responses:
Response 1 - @dahlihwa good morning 😄...
Response 2- @dahlihwa Good morning em♡

Topic 5: Missing and Nostalgia:
Post : i hope yall miss me and my dumb tweets...

Responses:
Response 1 - @merlotmv dont worry im back again (until the next concert)...
Response 2 - @haonslut i will miss u sm :(...
Response 3- @95MINSIK ill be back...
Response 4- @emmys_archive so true...
Response 5- @justerithings true same...
Response 6- @haonslut yes i miss you and your tweets...
Response 7- @haonslut i will miss you today tomorrow and forever...
Response 8- @haonslut I always miss us...
Response 9- @tyyunseo im gonna be back soon

Future Work

- We can apply pre-trained models (Transfer Learning) to enhance performance.
- We can validate the dataset by manually labeling a small subset to ensure accuracy and quality.
- We can apply the Generative adversarial Networks (GANs) to increase the size of the dataset.
- For Bertopic, we can fine tune more to achieve better results.
- We can collect more data from some source.
- We can use any other topic model like top2Vec, lda2Vec etc. for topic modelling.

Conclusion

- **Best Model and Key Findings**
  Non-Negative Matrix Factorization (**NMF**) emerged as the optimal topic model for our analysis, achieving a coherence score of 0.656. Our investigation of Twitter conversations using NMF and PyLDAvis revealed a strong presence of social support within K-Pop online communities**. Emotional support** constituted the most prevalent form, accounting for **46.6% of tokens.**

  While LDA and BERTopic exhibited average performance with coherence scores of 0.599 and 0.59 respectively, and LSA demonstrated a lower coherence score of 0.41, these models still contributed to the overall understanding of the topic space. Notably, the performance of BERTopic using Voyage AI embeddings was comparable to that of TF-IDF embeddings, indicating potential for further exploration of different embedding techniques.

- **Consistency in Findings**
  All four topic models consistently identified core themes related to social support. Emotional support was primarily linked to discussions of personal challenges, while appraisal support was evident in opinion-sharing and achievement-focused conversations. Informational support surfaced in academic-related discussions and merchandise purchases, and instrumental support was observed in conversations about purchasing processes and events.

- **Model Combination and Limitations**
  The combined application of topic modelling and large language models enabled the comprehensive identification of social support within the analysed text. However, challenges were encountered in effectively detecting social emotional support within Reddit data compared to Twitter data.

  Overall, this study demonstrates the efficacy of combining topic modelling and LLMs to uncover the multifaceted nature of social support within K-Pop fandoms on social media.

References

[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] https://medium.com/@iqra.bismi/topic-modelling-using-lda-fe81a2a806e0

[6] https://cran.r-project.org/web/packages/nnTensor/vignettes/nnTensor-1.html#:~:text=The%20reconstruction%20error%20(%20RecError%20)%20and,calculation%20is%20converging%20or%20not.&text=The%20product%20of%20U%20and,be%20well%2Drecovered%20by%20NMF%20.

[7] https://towardsdatascience.com/latent-semantic-analysis-intuition-math-implementation-a194aff870f8

[8] O'Callaghan, Joe. "c_h: Topic Coherence Explained." Towards Data Science, Medium, 26 Oct. 2021, towardsdatascience.com/c_h-topic-coherence-explained-fc70e2a85227. Accessed 12 Aug. 2024. 22

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

Additional Files:

Poster: Final Poster.pptx
Gitlab link: https://gitlab.socs.uoguelph.ca/sscott15/social-support-in-kpop-fandoms/-/tree/Project?ref_type=heads

My GitHub: https://github.com/AditiSatsangi/Globalizing-K-Pop-Project-Analysing-Social-Support-using-Topic-Modelling-and-LLMs

Link for Data Pre-Processing:
https://colab.research.google.com/drive/1QzXUDGwNHdqBfs7UoVT3jyr_VwWn9qjO?usp=sharing

Link for code related to Topic modelling: LDA, NMF, LSA:
https://colab.research.google.com/drive/1i-hdc9CkKfKcLECtmv93rHjU0ug8_epW?usp=sharing

Link for code related to BERTopic:
https://colab.research.google.com/drive/11vGZNBQgN5jGXF8wn-5v9ymvfSuhTg9s?usp=sharing

Link for the GitHub Data (related to paper) implementation (Topic Modelling):
https://colab.research.google.com/drive/1Xn6SUDmqlrJj98Blrw-RLgfC6kzaE9FM?usp=sharing

Deployed link: https://aditisatsangi.github.io/Globalizing-K-Pop-Project-Analysing-Social-Support-using-Topic-Modelling-and-LLMs/#topic=0&lambda=1&term= (pyLDAVis visualization)