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

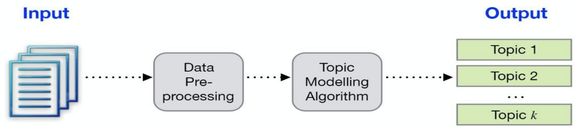
**Aditi Satsangi, Dr. Stacey Scott** and **Dr. Ritu Chaturvedi, School of Computer Science, University of Guelph**

## Introduction

- Social media platforms like Twitter and Reddit have become vital platforms for online discussions particularly within K-pop fandoms. We extracted the data from twitter and reddit both.
- This research uses **topic modelling** and **natural language processing techniques** to understand emotions, opinions and social support within these fandoms [1]. LLMs are used to represent results of the topic models.

## Background

- **K-pop** is a form of popular music originating in South Korea as part of South Korean culture.
- **Topic Modelling** is a statistical technique which is used to discover latent topics that exist within a collection of documents.
- **Large Language Models (LLMs)**: AI models trained on vast text data to understand and generate human language, used in content creation.




## Data Description

Twitter:

*Post 1*: 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

*Response 1*: i love u and i hope things get better for you soon

*Response 2*: love u more than u know my jaspie 💗 <gif>

*Response 3*: we love you too pup 💗

*Response 4*: 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 <gif>

Reddit:

*Post 1:* Hi everyone! I joined an Sf9 album group order and a fan call was up to win and I won!! I was told I can pick which member I can talk to and I'm planning to talk to Zuho. I never did a kpop fan call and I'm just wondering how does it work? Do y'all have experiences or advice? 😊

*Response 1:* I have no idea but I'm so happy for you! Have a great time 🎁🖤what is this pc from? i got it as a random freebie from kpopusa

*Response 2:* Omg I'm so happy for you! How was your experiences with other members too?

## Methodology

**1. Data Collection**
- Collected Reddit data using PRAW (web crawler)
- Collected Twitter data related to K-pop fandoms using Twitter API

**2. Data Pre-Processing**
- Linked conversations using conversation IDs
- Cleaned data: removed emojis, hashtags, special characters, and fandom names; translated non-English text to English
- Converted text to lower case
- Performed tokenization, stop word removal and lemmatization
- Applied Keyword filtering

**3. Text Representation**
- Term Frequency-Inverse Document Frequency (TF-IDF) used to represent text
- This transformed raw text into numerical format for analysis

**4. Topic Modelling**
- Applied three topic models for extracting topics from the text data:
  - Latent Dirichlet Allocation (LDA)
  - Non-Negative Matrix Factorization (NMF)
  - Latent Semantic Analysis (LSA)

**5. Evaluation**
- Calculated coherence score for comparison of different topic models
- Used metrics perplexity (LDA), reconstruction error (NMF), and singular values (LSA)
- Visualized results with pyLDAvis for NMF (see Fig. 1).

**6. Representation using Large Language Models (LLMs)**
- Employed LLMs for result interpretation:
  - KeyBERT: Keyword extraction for each topic
  - BART: Topic Summarization
  - GPT: Topic labelling

*Scan this code to access this poster and citations and suggestions/ feedback form*

## Results

- We identified five topics related to social support. Among these, emotional support is the most predominant in the text.

*Table 1: The Coherence Scores for different Topic models used*

| Topic Model | Coherence Score |
|---|---|
| LDA | 0.599 |
| NMF | 0.656 |
| LSA | 0.416 |

- Model comparisons: NMF performed best, achieving the highest coherence score of 0.656.

*Table 2: 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... |
| 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... |

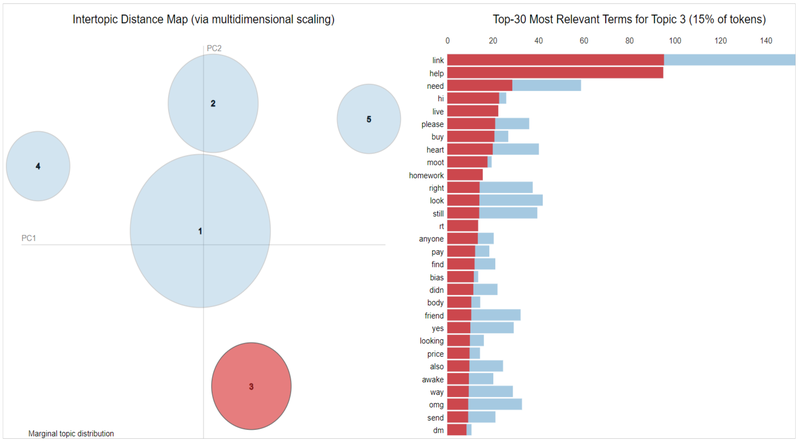



*Fig 1: Visualization of clusters obtained from NMF model using pyLDAVis*

## 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.

## References

[1] Egger, Roman, and Joanne Yu. "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts." *Frontiers in sociology* 7 (2022).