From LLM-Enhanced Topic Modeling to Applications: A Case Study on Tracking K-pop Social Footprint

Aylin Fathi [1], Juwon Shin [1], Kyungsik Han [1,2]

[1] Department of Data Science, Hanyang University, Seoul, Republic of Korea
[2] Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea

Research Motivations

Limitations & Research Gap

- **Unsupervised** topic modeling often requires subjective **manual** naming, while **LLM-only labeling** at scale can be **costly** and **inconsistent**.
- Many **topic studies** remain largely **static**, offering **weak** support for **identifying trends** in **evolving communities**.

Importance of Analysis

- **Lyrics**: Lyric theme analysis helps map the dominant narratives in K-pop songs.
- **Gender bias**: Bias-focused analysis can reveal subtle gendered assumptions embedded in popular lyrical language.
- **Reddit trends**: Trend analysis highlights how public discussion priorities change, tracking both consistent and emerging perspectives.

Importance of Platform

- **External knowledge retrieval** reduces **hallucination** in LLMs and improves **credibility**.
- Topic-indexed outputs give listeners **direct control** over their interests, rather than relying on **algorithmic** recommendations.
- **Connecting** research **analysis** to real-world **application**.

We propose a lyric-centered analytical **framework** that extracts **lyric themes** and enables a **topic-indexed database** for **music recommendation**, while also **quantifying** embedding-based **gender bias** and tracking Reddit discussion **trends**.

Methodology

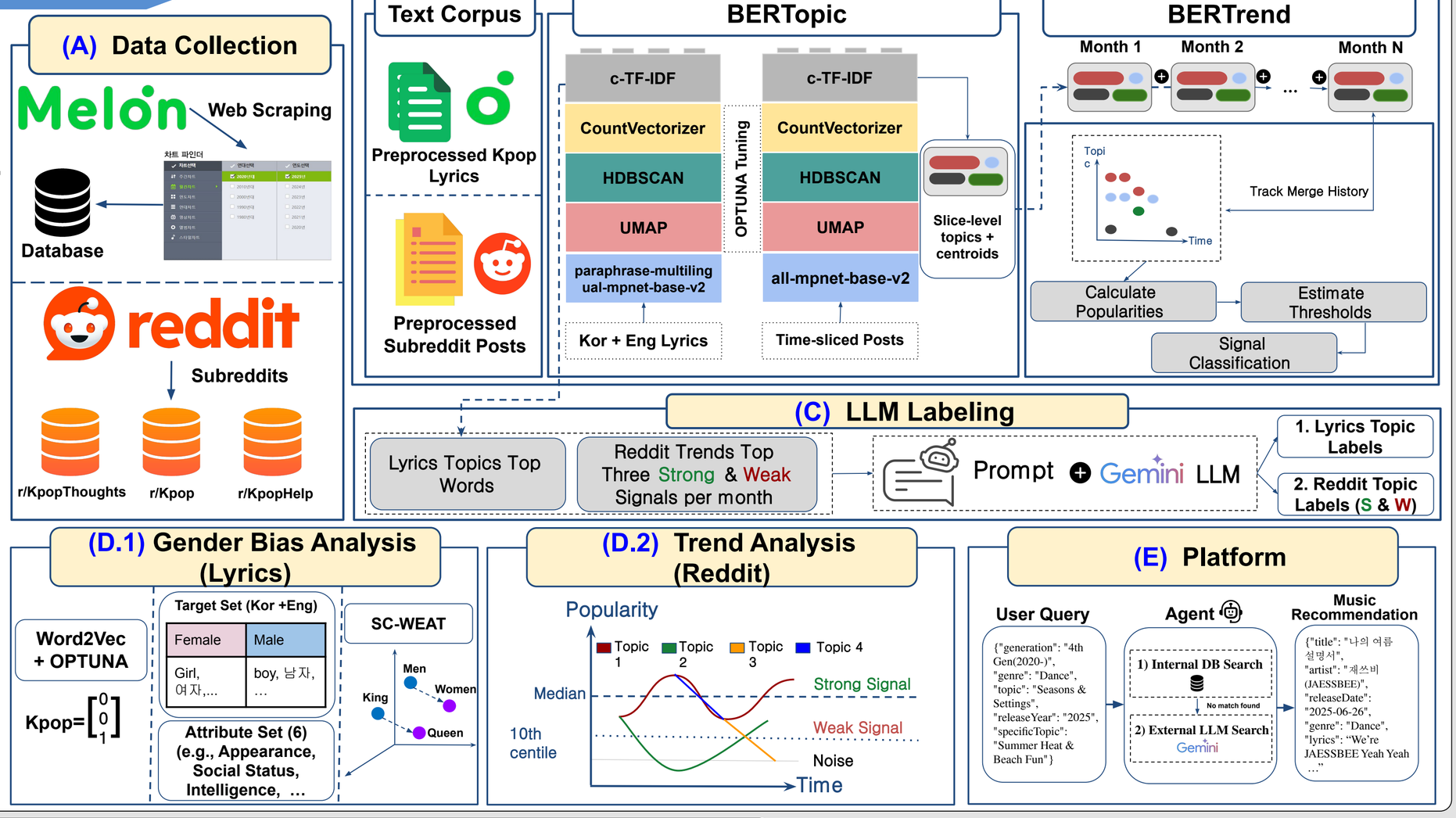
(A) Collected Melon **lyrics**, metadata and Reddit K-pop **posts** to **capture song themes and fan discussions**.

(B) Optuna-tuned **BERTopic** with **multilingual, all-mpnet embeddings.** BERTrend merges monthly **slice-topics** into **global topics** using **cosine similarity** of **topic-centroid** embeddings.

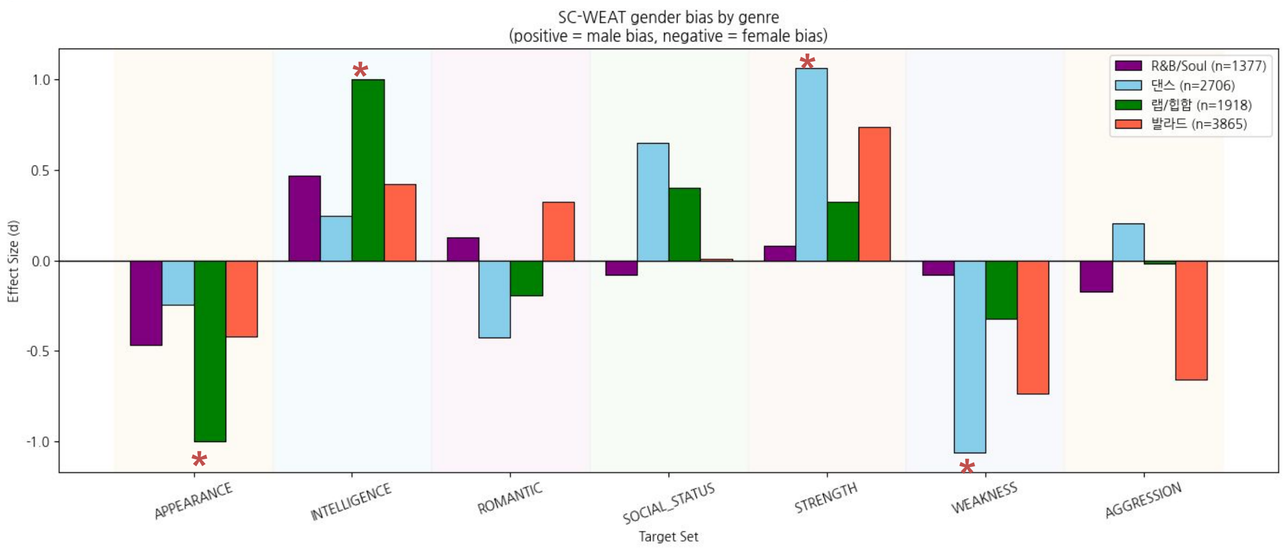
(C) Used Gemini to generate **structured labels** from **topic top words** in **lyrics & posts** (strong & weak Reddit signals).

(D) Quantified gender bias in lyrics by **Word2Vec** and applying **SC-WEAT** with **male, female target** sets and multiple **attribute sets**. Analyzed BERTrend **popularity signals** to identify **emerging vs. consistent** themes.

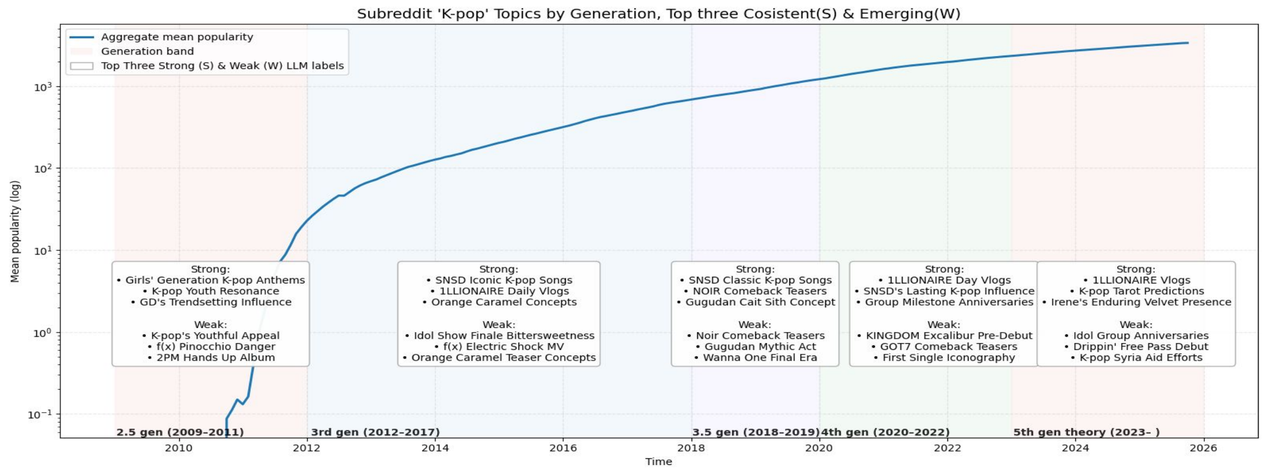
(E) A **two step ReAct** process. It first performs an **Internal DB Search**, and only when no matches are found, proceeds to **External LLM Search**.



Gender Bias & Trend Results



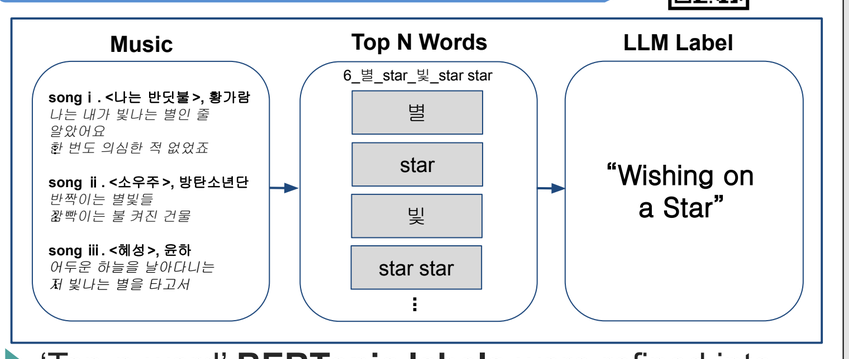
- **Male bias** towards **intelligence, strength**, & almost **social status**
- **Female bias** for **appearance, weakness**
- Highlights **genre-specific shaping** in Kpop.



- **Consistent (Strong)** Shift from famous artist discussions to broader fandom culture (e.g.,vlogs, K-pop tarot.
- **Emerging (Weak)** Signals like "K-pop Syria aid" highlight occasional real-world mobilization forums.

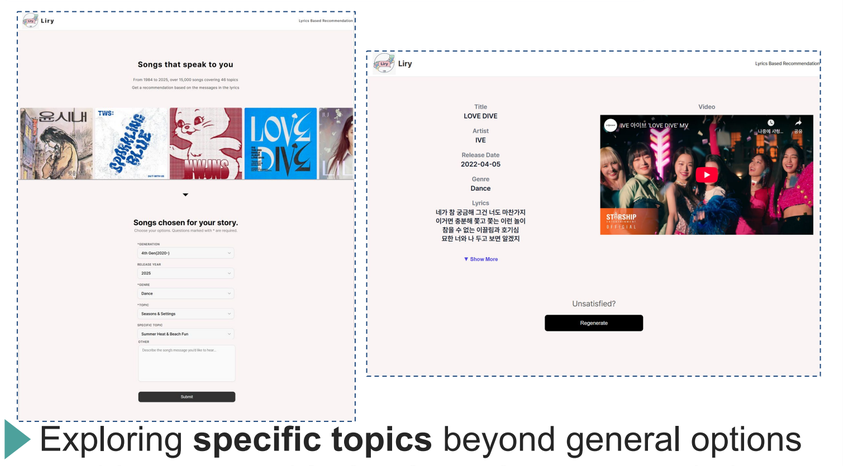
LLM Topic Labeling

Topic Labeling (2024-only)



- 'Top-n-word' **BERTopic labels** were refined into **objective labels** using Gemini model.

Platform (Liry) Results



- Exploring **specific topics** beyond general options enables targeted **lyrics-based** recommendations.

Conclusion & Future Work

- We demonstrated that K-pop lyrics show measurable **gender stereotypes** and that **large-scale** public discussions are **identifiable** over time, enabling a **quantitative** view of K-pop's **social footprint** beyond relying only on a single source of music.
- Through **topic modeling**, we identified **themes** hidden in lyrics, organized them into a database that **improves reliability**, **reduces hallucination** in LLMs, and provides **topic-based music recommendation**.
- We expect to enhance this work by **expanding the lyrics corpus**, testing **multiple LLMs** labeling, and integrate **more features** to our platform.