

# Recommendation of Korean-Pop Bands using Topic Modelling Algorithm and Myers-Briggs Type Indicator

MSc Research Project  
Data Analytics

Nurul Hanis Binti Ibrahim  
Student ID: x20246862

School of Computing  
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland  
Project Submission Sheet  
School of Computing



<b>Student Name:</b>	Nurul Hanis Binti Ibrahim
<b>Student ID:</b>	x20246862
<b>Programme:</b>	Data Analytics
<b>Year:</b>	2022
<b>Module:</b>	MSc Research Project
<b>Supervisor:</b>	Dr. Catherine Mulwa
<b>Submission Due Date:</b>	15/12/2022
<b>Project Title:</b>	Recommendation of Korean-Pop Bands using Topic Modelling Algorithm and Myers-Briggs Type Indicator
<b>Word Count:</b>	3381
<b>Page Count:</b>	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

<b>Signature:</b>	Nurul Hanis Binti Ibrahim
<b>Date:</b>	30th January 2023

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission</b> , to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project</b> , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Recommendation of Korean-Pop Bands using Topic Modelling Algorithm and Myers-Briggs Type Indicator

Nurul Hanis Binti Ibrahim  
x20246862

## Abstract

Enhancing the quality of information retrieval tasks comes a greater challenge for music streaming services. Particularly, in researching and developing a robust recommendation algorithm that fulfils the demand of users in providing music artist recommendations without being biased towards particular personality types as much as possible. In the light of this, this research focus on recommending the novel Korean-pop bands by utilising the Latent Dirichlet Allocation, a topic modelling algorithm and a personality framework, the Myers-Briggs Type Indicator. Practical experiments of the content-based recommendation algorithm shows that the combination of the aforementioned has successfully produce a variety of K-pop bands in accordance to two of the personality type listed in the framework.

## 1 Introduction

### 1.1 Background and Motivation

Korean-pop (K-pop) songs in digital format have reach billions of its listeners across the globe with the rapid advancement of technologies. This phenomenon leads to the soaring demand for unique K-pop band recommendations to its avid and active listeners respectively. Adjacently, pressuring music streaming services, such as Spotify, Apple Music and YouTube Music, in catering to the listeners' distinctive personalities. Generally, a domain-centric recommender engine, such as the K-pop bands in this scope of research, operates when the algorithms produce a match between the users' representation and the band's representation to find the relevant band to the users. Currently in practice, the recommender algorithms and the parameters adopted by the music streaming services are lacking in providing the precise band recommendations for diverse personalities of individuals. The impreciseness of the system stems from the type of the parameters, in which are the personality framework that expresses users' representation entity and the algorithm that manifests the band's representation entity. These entities that making up the system is contributing to the biasedness towards the uncommon personality groups. In solving the practical problem, this work proposed to recommend a set of K-pop bands, for a given personality type based on the Myers-Briggs Personality Indicator (MBTI), using Latent Dirichlet Allocation (LDA) algorithm, that is the subset of Natural Language Processing (NLP) techniques, embed in the content-based recommender algorithm.

## 1.2 Research Question, Objectives and Document Roadmap

Analysing users' explicit feedback in textual format, such as comments on specific item, requires the ability of topic modelling algorithms for enhancing information retrieval tasks. Simply, explicit feedback provided by users is a useful feature to validate the reliability of user's preference (Mandal & Maiti 2018) of different MBTI types. Furthermore, the rationale of the topic modelling algorithms in analysing the explicit feedback is to improve the relatedness of search results (Rajapaksha & Silva 2019). Hence, to demonstrate their suitability in increasing the quality of content-based recommendation algorithm, the following research question is presented for this project:

*To what extent can Korean-pop bands recommended to users using Latent Dirichlet Allocation algorithm with combination of Myers-Briggs Type Indicator to reduce personality bias?*

To solve the research question, a set of objectives are outlined and implemented:

**Objective 1:** Critical review of the past research works pertaining to the scope of the project.

**Objective 2:** Design and creation of an online questionnaire for data collection.

**Objective 2(a):** Preparing the collected data for subsequent implementations

**Objective 3:** Implementation of Latent Dirichlet Allocation algorithm.

**Objective 4:** Implementation of content-based recommendation algorithm that uses Myers-Briggs Type Indicator as input to generate recommendations

**Objective 5:** Evaluation, Results, and Discussion of Latent Dirichlet Allocation algorithm. (**Objective 3**)

**Objective 6:** Evaluation, Results and Discussion of content-based recommendation algorithm (**Objective 4**).

**Objective 6(a):** Experiment: Testing input 'ENFP' of MBTI type.

**Objective 6(b):** Experiment: Testing input 'ISTJ' of MBTI type.

The structure of this technical report extends from Critical Review of Past Research Works to Research Methodology and Design Specifications. The remaining part of the report covers Implementation, Evaluation, Results and Discussions, and Conclusions and Future Work.

## 2 Critical Review of Past Research Works

In subsection 2.1, the author critically reviewed the application of personality traits in recommender systems developed by past researchers. This section extends to subsection 2.2, whereby the author critically review of the application of topic modelling algorithm in content-based recommender system, and subsection 2.3, briefly discussed about the current state of research relating to this project domain. In the final part, the prospect of bringing together the personality traits and topic modelling algorithm in reducing personality bias is justified.

### 2.1 Critical Review of the Application of Personality Traits in Recommender Systems

The author has reviewed critically of the past works that have used personality background to develop a more personalisation method, the problem with the Five Factor Model personality framework, and justifies the role of MBTI as the main personality framework of this research.

#### 2.1.1 Using Personality Background to Develop a More Personalisation Method

Some studies have included users' personality traits in developing recommender systems. Particularly, in the music domain. A team of researchers (Tkalčić et al. 2015) have discovered that the personality for each individual says a lot about their likeability towards a specific music genre. This means that the level of consumption of a music genre is higher for different personality types. This dynamic relationship provides the fundamental evidence that the personality entity can be associated with specific multimedia contents. Concerning about the evaluation, a study (Paiva et al. 2017) have achieved greater than 75 % user satisfaction rate of the given recommended items based on their personality profiles. Thus, justifying that using personality entity enables the development for a more personalised system.

#### 2.1.2 The Problem with the Five Factor Model Personality Framework

Most of the previous works in modelling user personality have used the Five Factor Model (FFM) framework. However, the efficacy of the framework could have done more in terms of modelling users' personality. This comes as it decreases the ability of a recommender system to produce concrete evaluation outcomes (Guntuku et al. 2018). It is worth to note that the FFM only manifest itself into five dimensions, which are openness, conscientiousness, extroversion, agreeableness, and neuroticism. Hence, by profiling users in such restrictive approach, the system will overlook the deeper insights of 'who likes what and why'. In accordance to that, applying this homogeneous and small-scale personality grouping method has proven to produce less accurate results as reported in another study (Szymdt 2021). Overall, it is concerning of the gaps that FFM have left based on the evidences aforementioned. Therefore, it is imperative to considering an alternative personality framework that has a wider inventory. In which, able to capture subtle personality traits and the popular ones synonymously.

### 2.1.3 Myers-Briggs Type Indicator as the Main Personality Framework of this Research

Indeed, it is difficult to predict human behaviour as it is based on many dimensions ambiguously (Guntuku et al. 2018). As well of having to undergo the taxing task of gathering user experience through the means of questionnaires (Szmydt 2021). Regardless, MBTI is currently the best possible choice of personality framework to solve the impracticalities in the previous discussion. MBTI is a personality inventory that constitutes a total of 16 personality types from the combination of two poles in four different categories of an individual's propensity, which are, energy, perceiving, judging, and orientation, that is expressed into four letters. For instance, if a person is inclined to Extroversion (E), Intuition (I), Thinking (T), Perceiving (P), the person has the ENTP personality type (Woods RA 2022). As an illustration, Figure 3 shows the four dimensions of MBTI (Nagahi et al. 2020). Whereas, in Figure 2 shows the combination of the 16 MBTI personality types (Park 2017). As per the explanation given, MBTI is easier to be interpreted by the community that is not as expert as scientists and psychologists (Piedboeuf et al. 2019). Therefore, using MBTI as the main personality framework in this research will improve the robustness in measuring users' perception of the recommending content.



Figure 1:  
*The four dimensions of MBTI.*

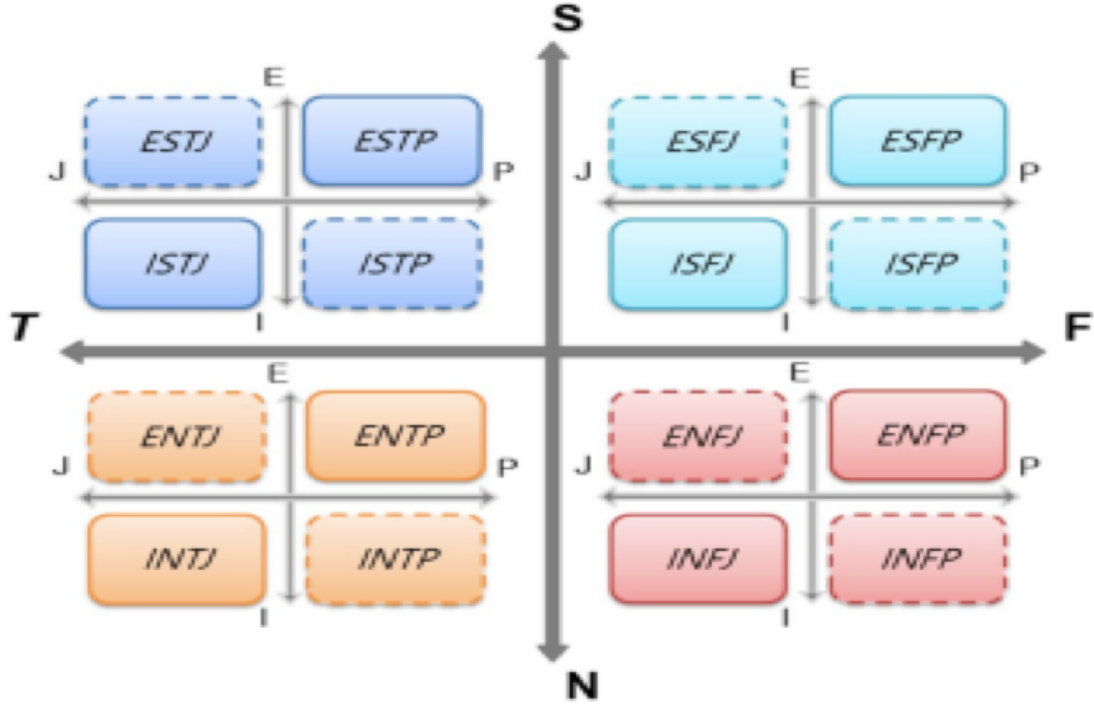


Figure 2:  
*The 16 MBTI personality types*

## 2.2 Critical Review of the Application of Topic Modelling Algorithm in Content-Based Recommender Systems

In this subsection, the author has reviewed critically of the past works that have utilised topic modelling for sentiment analysis of social media data.

### 2.2.1 Utilising Topic Modelling for Sentiment Analysis of Social Media Data

Topic modelling has become an increasingly important algorithm to understand the users' opinion towards certain items such as artists' reviews from social media platforms. In the sense that the algorithm is able to identify important topics quickly and coherently. A study confirmed this notion by successfully learning the thoughts, attitudes, feelings of individuals towards the COVID-19 pandemic by employing a topic modelling algorithm to a social media data (Abd-Alrazaq et al. 2020). Simply, topic modelling is able to extend its purpose of extracting sentiments. Albeit, this algorithm fits into short texts only as discovered in two studies by Xiong et al. (2018) and Xu et al. (2018) respectively. It is important to note that user-generated content on social media, such as textual comments, is not lengthy. Hence, this type of constraint could better demonstrating the algorithm's fitness in the domain of processing social media data.

## 2.3 The Personality Bias Problem in Content-based Recommender System

To the best of author's knowledge, no research has been done to address the personality bias problem in content-based recommender systems specifically. This is because prom-

inent researches (Li et al. 2021), (Ge et al. 2021) involving fairness in recommendation mainly focused on the collaborative filtering recommender system. Hence, this is the first attempt to reduce personality bias, which is a concern raised in the fairness aspect, in a different type of recommender system.

The gap raised in the aforementioned is overlooked. This has given the opportunity for the author to contribute to the body of Data Analytic knowledge. In a way of bringing MBTI, to capture user’s perception of the items, and topic modelling algorithm in enhancing the sentiment analysis of that perception to reduce personality bias in a content-based recommender system. Overall, the critical review of all the literature in this section has informed the author of the current state of research and its gaps respectively. That is, within the scope of this project.

Overall, this chapter has achieved Objective 1. That is, contributed to the understanding of the current state of research and its identified gaps.

### 3 Research Methodology and Design Specification

#### 3.1 Research Methodology

Following a modified Cross-Industry Standard Process for Data Mining (CRISP-DM) as the research methodology is essential in the development of the band recommender algorithm. The iterative approach of CRISP-DM outweighs the linear approach of Knowledge Discovery in Databases (KDD). In the sense of being able to avoid any discrepancies from each cycle from overlooking the business objective. Thus, CRISP-DM saves the time, cost, and resources in running the project prior to its deployment. The modified methodology in Figure 3 is also within the Business Logic Tier in Figure 4, illustrating the interplay between itself and the design specification.

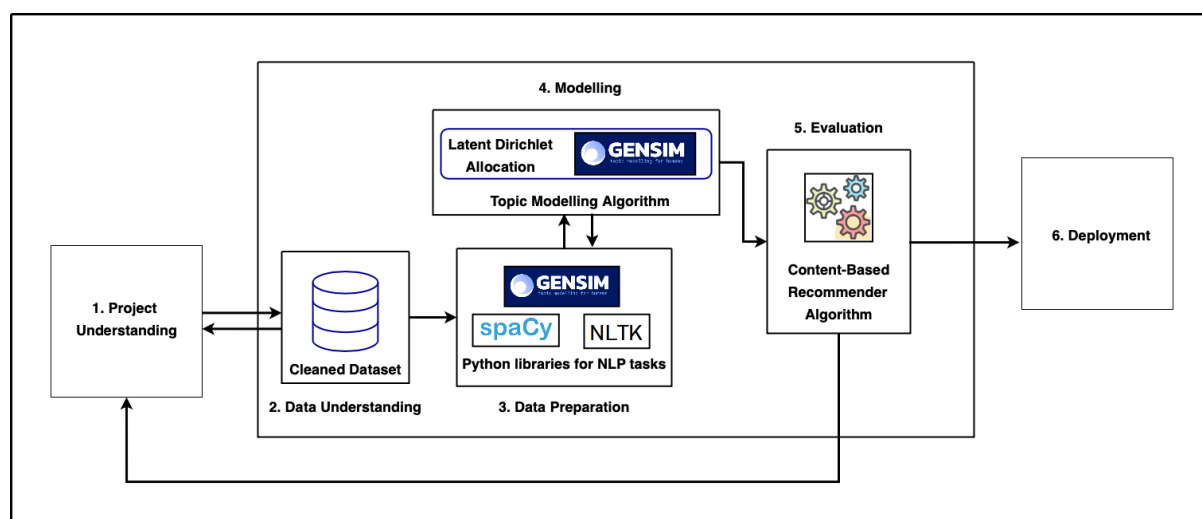


Figure 3:  
Korean-Band Recommender Methodology Approach (Business Logic Tier)



The above diagram is explained below:

1. **Project Understanding:** The project goals and its subsequent planning are formulated based on the business problem. Provided that this research aimed to build a K-pop band recommendation engine, thus, the engine shall recommend the top 3 K-pop bands based on the input of the users' MBTI personality type.
2. **Data Understanding:** The dataset is collected by distributing a questionnaire in Google Forms to online users anonymously through Reddit, a social media platform. The questionnaire's content is shown in Table 1 below. From Question 4 until Question 6, the author selected 30 popular K-pop bands. A total of 600 questionnaire replies has been obtained.
3. **Data preparation:** This phase pertains to the data cleaning and data transformation prior to modelling. In data cleaning, specific attributes are extracted from the collected questionnaire replies in the previous step as shown in Table 2. The attributes names are changed into simpler words for ease of the data modelling. The removal of null and duplicated values, special characters, numerical characters, newline characters and whitespace for every extracted attribute is also part of the data cleaning process. In arriving at the final data format for the modelling process, the 'MBTI' column is split into 16 separate DataFrames as per the original 16 types of MBTI. For each of the Data Frames, the Comments column is created as per the combination of the columns Comment\_Band\_1, Comment\_Band\_2, and Comment\_Band\_3. The 3 columns are removed afterwards. The DataFrames are stacked on top of each other, resulting into a single DataFrame that is cleaned for the processing phase. Meanwhile, in the aspect of data transformation, it is important to note that the recommender algorithm deals with textual data, which is the user's comments for their K-pop bands of interest, in generating the appropriate set of band recommendations. By this notion, this research employed Python libraries for NLP tasks such as Gensim, SpaCY, and NLTK on those textual data before applying into the LDA algorithm. The tasks of NLP consist of tokenization, stop words removal, data lemmatization, creating a dictionary and convert it into a Bag of Words (BoW) format. The explanation of each tasks is tabulated in Table Table 3.
4. **Modelling:** The data is prepared to be modelled by LDA to extract topics within the comments of the users. LDA is assumed that the documents, are made up from a mixture of topics chosen on the basis of Dirichlet distribution over a fixed set of topics. Then, the topics generate words assigned with certain probability scores on the basis of the multinomial distribution of each topic (Yang 2015). Hence, the aforementioned process are mathematically defined<sup>2</sup> in Figure 4. In the context of this project, LDA is able to identify hidden topics for each comment (McAllister et al. 2019) given by users belonging to the 16 different types of MBTI. The topics are represented by word probabilities (Jelodar et al. 2018). Thus, the highest probabilities of words per topic says a lot about the users' sentiments towards their K-pop bands of interest. This follows an example in another domain<sup>3</sup> illustrated in Figure 5.

---

<sup>2</sup><https://datascienceplus.com/topic-modeling-and-latent-dirichlet-allocation-lda/>

<sup>3</sup><https://towardsdatascience.com/dimensionality-reduction-with-latent-dirichlet-allocation-8d73c586738c>

5. **Evaluation:** The LDA model and the recommender algorithm is validated and discussed in this phase. This is demonstrated by partitioning the comments into two parts and implementing them into an analysis. That is, by computing the LDA model transformation using cosine similarity<sup>5</sup>. Its formula in Figure 6 illustrates in a manner that two vectors of the divided comments, denoted by A, representing the target comments' word vector and B, representing the compared comments' word vector. The numerator of the formula depicts the shared word vectors between A and B. Meanwhile, the denominator refers to the number of words in both vectors. Whereas, the recommender algorithm is evaluated by experimentation. In the way of observing the resultant output from the algorithm inputs. The experiment results will mapped back to the project understanding stage. If the results match with the goals then deployment can be proceeded.
6. **Deployment:** The final part of the modified methodology is embedding the algorithm to a Graphical user Interface (GUI) application using Anvil, a drag-and-drop web app builder that is written in the Python programming language. The Figure 8 showcase the band recommender in the form of the web application.

Table 1: Questionnaire's content

No.	Questions	Choice/Input
1	What is your MBTI Type?	ISTJ, ISTP, ISFJ, ISFP, INTJ, INTP, INFJ, INFP, ESTJ, ESTP, ESFJ, ESFP, ENTJ, ENTP, ENFJ, ENFP
2	Which age group you belonging to?	18 - 23 years old, 24 - 29 years old, 30 - 35 years old, Above 35 years old
3	Based on the list below, write your top 3 K-pop bands.	SEVENTEEN, ITZY, ASTRO, (G)-IDLE, ATEEZ, Enhypen, KARD, STAYC, aespa, Golden Child, Treasure, NCT 127, Blackpink, BTS, Twice, The Rose, Day6, TOMORROW X TOGETHER, Stray Kids, IZ*ONE, N.FLYING, Brave Girls, GFRIEND, Pentagon, Dreamcatcher, Weki Meki, OH MY GIRL, MONSTA X, IKON, MAMAMOO
4	Based on your previous selection, what do you like about the Group 1?	(User's own input)
5	Based on your previous selection, what do you like about the Group 2?	(User's own input)
6	Based on your previous selection, what do you like about the Group 3?	(User's own input)

<sup>5</sup><https://sites.temple.edu/tudsc/2017/03/30/measuring-similarity-between-texts-in-python/>

Table 2: Extracted Data and Name Changing Convention

Extracted Attributes	Name Changes
What is your MBTI Type?	MBTI
Based on the list below, write your top 3 K-pop bands.	Band
Based on your previous selection, what do you like about the Group 1?	Comment_Band_1
Based on your previous selection, what do you like about the Group 2?	Comment_Band_2
Based on your previous selection, what do you like about the Group 3?	Comment_Band_3

$$P(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \alpha, \beta) = \prod_{i=1}^M P(\theta_j; \alpha) \prod_{i=1}^K P(\phi; \beta) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \phi z_{j,t})$$

Figure 4:

*The Mathematical Model of Latent Dirichlet Allocation*

$\alpha, \beta = \text{Dirichlet distributions}$   
 $\theta, \phi = \text{Multinomial distributions}$   
 $Z = \text{Topic vectors of all words in all documents}$   
 $W = \text{Vectors with all words in all documents}$   
 $M = \text{Number of documents}$   
 $K = \text{Number of topics}$

## With Latent Variables

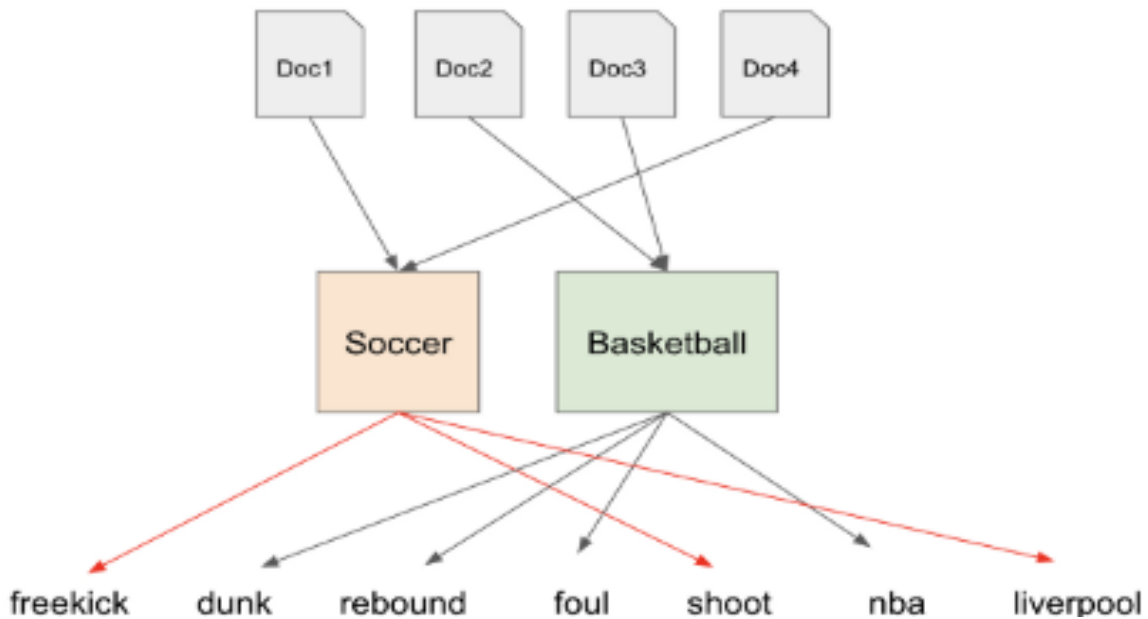


Figure 5:

*Latent variables in LDA*

Table 3: Further explanation of Natural Language Processing Tasks

Natural Language Processing Tasks	Explanation
Tokenization	The texts in the Comments column are split into single words or tokens, using <code>simple_preprocess</code> function imported from the Python's Gensim library.
Stopwords	The tokens are removed of its stopwords such as, a, an, the, for focusing on important words and reducing number of features. That is, helping in the optimization for the modelling phase.
Bigrams	It is important to note that some words in English could frequently occur together. For instance, 'high school' or 'best performance'. Hence, the author identified such pair that will help in sentiment analysis of the Comments column. Such pair are known as bigrams. Gensim is the provider of the bigram function in generating these pairs.
Lemmatization	The application of lemmatization involves using a dictionary to return the words into its root form. For instance, 'better' is derived from 'good', therefore lemmatization process would reduce the words into 'good'. This task is suitable in learning the texts meaningfully.
Dictionary	Prior to converting into Bag of Words, a dictionary is created. That is, the words in the corpus mapped with their integer representations. The dictionary is filtered by eliminating the words that appear in more than 10% of the corpus.
Bag of Words	As the final task, the corpus is converted into Bag of Words (BoW) format. The arbitrary nature of texts is converted into vectors of a fixed length. In this way, text data can be represented when modelling it with the LDA algorithm.

$$\cos \theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

Figure 6:  
*Cosine Similarity Formula*

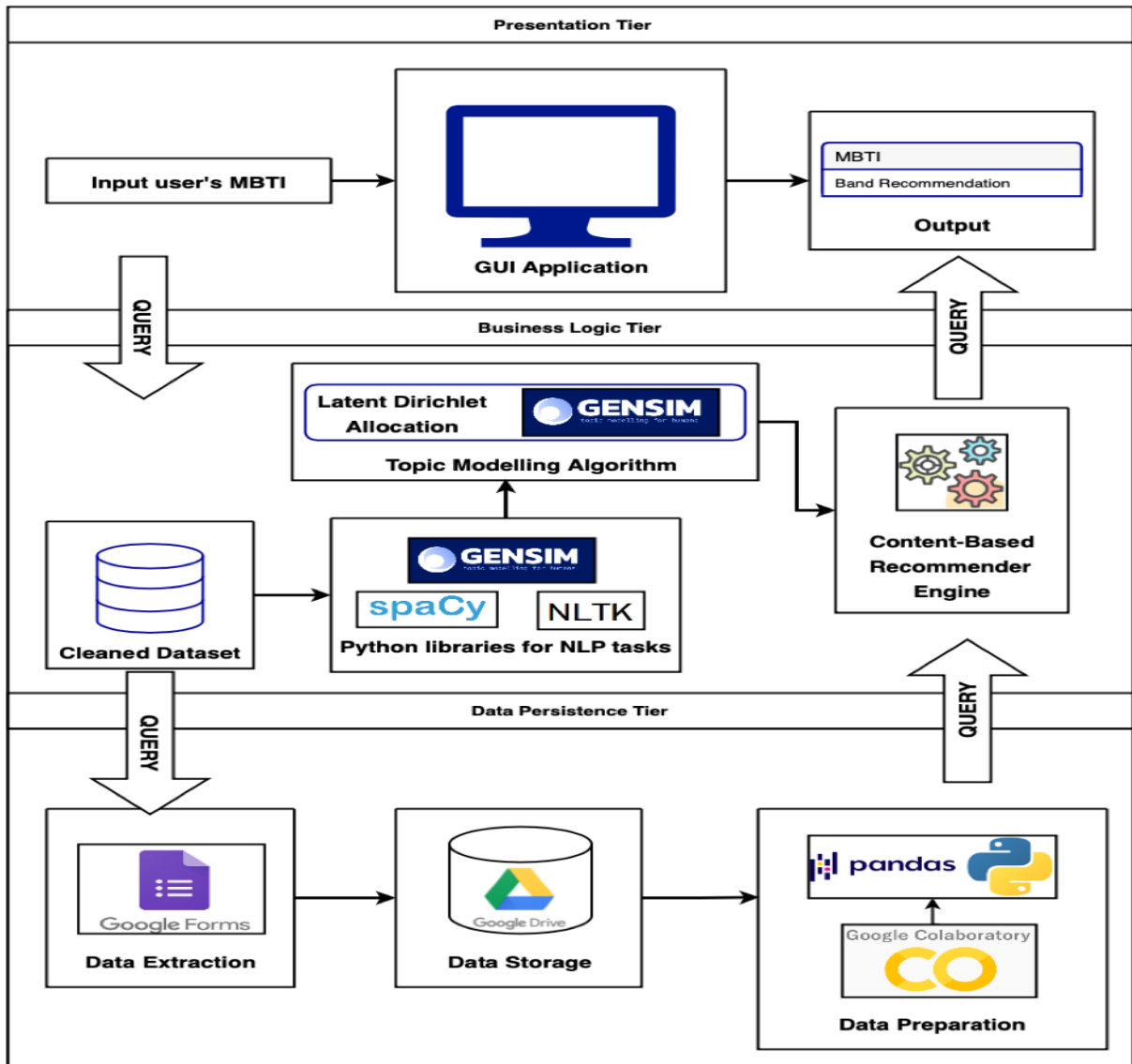


Figure 7:  
*Korean-band Recommender 3-Tier Design Architecture*

### 3.2 Design Specification

This research used the 3-Tier Design Decision as it employed cloud technologies. The interplay between the modified methodology and design is illustrated in Figure 7.

1. **Data Persistence Tier:** Data extraction and data storage are executed prior to preparing data. That is, to be forwarded to the Business Logic Tier. The raw data is extracted and stored using Google Cloud tools, which are Google Forms and Google Drive respectively. This is followed by importing and into Google Colaboratory using Pandas, one of the vast Python libraries.
2. **Business Logic Tier:** The purpose of this tier is to process and analysed data in achieving Objective 2.
3. **Presentation Tier:** In the final tier of the Design Decision diagram, the recom-

mender algorithm that is embedded into a Graphical User Interface (GUI) application written in Python is evaluated by inputting the MBTI. Subsequently, the recommendation output is generated for each input.

Overall, the aforementioned process has achieved Objective 2 and Objective 2(a). In respect to Objective 2, the online questionnaire is designed and created as shown in Table 1. Whereas, in respect to Objective 2(a) the data underwent changes in Table 2 and underwent NLP tasks as described in Table 3.

## 4 Implementation

### 4.1 Implementation of Latent Dirichlet Allocation

Table 4: Probabilities of Words per Topic

Topic Number	Word Probabilities
5	('love': 0.045020062), ('music': 0.035682864), ('music_video': 0.030435821), ('watch': 0.025098918), ('concept': 0.024309983), ('vocal': 0.022627125), ('side': 0.016587285), ('fun': 0.015430902), ('make': 0.014448428), ('dance': 0.014319093)
13	('music', 0.03402871), ('song', 0.033145692), ('really', 0.023836821), ('love', 0.023588179), ('well', 0.017178629), ('make', 0.01553155), ('member', 0.015172503), ('also', 0.01488779), ('personality', 0.01304192), ('feel', 0.012682035)
27	('concept', 0.06954491), ('catchy_song', 0.05317573), ('great', 0.041507453), ('member', 0.033861138), ('unique', 0.030046783), ('group_chemistry', 0.028850557), ('choreography', 0.024892302), ('music', 0.02238582), ('good', 0.02018631), ('look', 0.019927306)

In the context of this project, LDA is able to identify hidden topics for each comment (McAllister et al. 2019) given by users belonging to the 16 different types of MBTI. The topics are represented by word probabilities (Jelodar et al. 2018). Thus, the highest probabilities of words per topic says a lot about the users' sentiments towards their K-pop bands of interest as shown in Table 4

Hence, Objective 3 has been achieved.

### 4.2 Implementation of Content-based Recommendation Algorithm that uses Myers-Briggs Type Indicator as Input to Generate Recommendations

#### 4.2.1 Content-Based Recommender Algorithm

The algorithm 4.2.1 below illustrates the flow of the band recommender.

---

**Algorithm 1** Band Recommender

---

**Require:** similarity Matrix between all articles in the LDA model  $\leftarrow index$

**Require:**  $recommendation\_scores = \emptyset$

```
1: Input:  $mbti$ 
2: for  $row \in Comments$  do
3:   if  $mbti \leftarrow row$  then
4:     for  $mbti \leftarrow m$  do
5:        $index[lda\_vectors] \leftarrow sims$ 
6:       for  $sim \in sims$  do
7:          $[row.name[1], sim[1]] \leftarrow recommendation\_score$ 
8:          $recommendation\_score \leftarrow recommendation\_scores$ 
9:          $sorted(recommendation\_scores) \leftarrow recommendation$ 
10:      return  $recommendation$ 
```

---

Algorithm 1: Band Recommender

The algorithm is embedded into a Graphical user Interface (GUI) application using Anvil, a drag-and-drop web app builder that is written in the Python programming language. The Figure 8 showcase the final form of the web application of the band recommender engine.

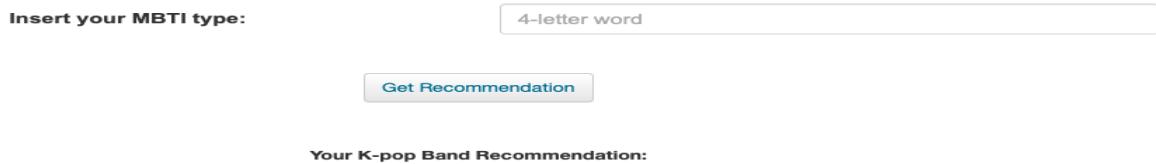


Figure 8:  
*GUI Application of the Band Recommender*

The front-end and the back-end is developed in the sense of able to display band recommendations with a click of button after input the desired MBTI type.

The detailed implementation of GUI application is discussed in the configuration manual. Hence, Objective 4 has been achieved.

## 5 Evaluation, Results and Discussions

The LDA model and the recommender algorithm are evaluated and discussed in this section.

### 5.1 Evaluation, Results and Discussion of Latent Dirichlet Allocation

In evaluating the LDA model, the similarity of documents is checked. The aim of the aforementioned is to retrieve documents that are similar and dissimilar semantically. It

is important to note that LDA is a topic modelling algorithm that is known for its unsupervised task. That is, having no expected label as a standard to compare the topics against unlike supervised learning. In the light of this, quality is measured in terms of the document similarity at the corpus level (Farkhod et al. 2021).

This is demonstrated by partitioning the comments into two parts and implementing them into an analysis. That is, by computing the LDA model transformation using cosine similarity<sup>5</sup>.

Table 5 shows the similarity score of topics. The intra-similarity score obtained suggests that the topics are moderately similar corresponding to the other topics in the document. Whereas, the low inter-similarity score obtained portray a low level of dissimilarity to other topics within 300 random parts of the document.

Table 5: Intra-similarity and Inter-similarity of the Comments

Intra-similarity: cosine similarity for corresponding parts of a document	0.57055
Inter-similarity: cosine similarity between random parts of a document	0.08270

Hence, Objective 5 has been achieved.

## 5.2 Evaluation, Results and Discussion of Content-based Recommender Algorithm

### 5.2.1 Experiment 1

Figure 9:  
*Experiment 1*

In the Figure 9 above, the recommender is shown to operate smoothly. ENFP, one of the 16 MBTI types is inserted as an input. The input is written in a sequence of 4-letter and is constricted by a match inside the recommender algorithm. As an observation, the resultant recommendation is displayed in the way that is readable. This means that the

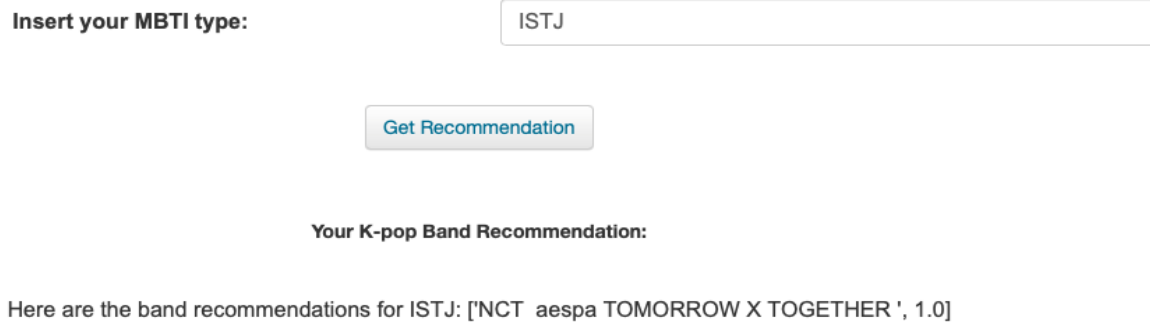
<sup>5</sup><https://sites.temple.edu/tudsc/2017/03/30/measuring-similarity-between-texts-in-python/>



input match the parameters of the recommender algorithm.

Hence, Objective 6(a) has been achieved.

### 5.2.2 Experiment 2



Insert your MBTI type:

[Get Recommendation](#)

**Your K-pop Band Recommendation:**

Here are the band recommendations for ISTJ: [NCT aespa TOMORROW X TOGETHER ', 1.0]

Figure 10:  
*Experiment 2*

Similarly, in the Figure 10 above, ISTJ, one of the 16 MBTI types is inserted as an input. The input is written in a sequence of 4-letter and is constricted by a match inside the recommender algorithm. As an observation, the resultant recommendation is displayed in the way that is readable. This means that the input match the parameters of the recommender algorithm.

Hence, Objective 6(b) has been achieved. Combining the achievement of Objective 6(a) and Objective 6(b) have brought out the achievement of Objective 6 overall.

## 6 Conclusion and Future Work

This research aimed to enhance the quality of information retrieval tasks of a content-based recommender algorithm by combining the MBTI, a personality framework that expresses user's representation entity, and LDA, in manifesting the explicit feedback of the Korean-pop band's as the entity's representation.

The main finding of this research is that topic modelling have demonstrated the capability in increasing the quality of content-based recommendation algorithm. In a way of capturing the relevant topics from user's sentiment to generate novel recommendation for each users' MBTI type. Thus, solving the problem of personality bias. Therefore, this has answered the research question. In the sense that K-pop bands can be recommended with topic modelling algorithm in tandem with the Myers-Briggs Type Indicator to reduce personality bias.

However, the author focus solely in the scope of Korean-pop bands due to the time constraint in completing this research. Hence, this research demonstrated to the best ability of the author within such time frame. Therefore, for future work, musical artist from

diverse genre such as Pop, Rock, and Jazz can be taken into account as well.

Moreover, the aforementioned solution would be a good fit within the domain of music streaming services. In the advent of Big Data, these services deals with a massive scale of data containing explicit feedback of users of certain musical bands. In which, facing the challenges of researching and developing a more personalised method to users. Therefore, this solution is able to contribute in addressing those challenges.

## Acknowledgements

The author would like to express gratitude to Dr Catherine Mulwa for her patience and determination in guiding and providing feedback to the author. This gratitude is also extended to the author’s beloved family and friends in providing mental, physical, spiritual and monetary support that helped the author in going through the ups and downs of producing this work.

## References

- Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M. & Shah, Z. (2020), ‘Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study’, *Journal of Medical Internet Research* **22**(4), e19016.  
**URL:** [https://doi.org/10.2196%2F19016](https://doi.org/10.2196/2F19016)
- Farkhod, A., Abdusalomov, A., Makhmudov, F. & Cho, Y. I. (2021), ‘LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model’, *Applied Sciences* **11**(23), 11091.  
**URL:** [https://doi.org/10.3390%2Fapp112311091](https://doi.org/10.3390/2Fapp112311091)
- Ge, Y., Liu, S., Gao, R., Xian, Y., Li, Y., Zhao, X., Pei, C., Sun, F., Ge, J., Ou, W. & Zhang, Y. (2021), Towards long-term fairness in recommendation, *in* ‘Proceedings of the 14th ACM International Conference on Web Search and Data Mining’, ACM.  
**URL:** [https://doi.org/10.1145%2F3437963.3441824](https://doi.org/10.1145/2F3437963.3441824)
- Guntuku, S. C., Zhou, J. T., Roy, S., Lin, W. & Tsang, I. W. (2018), “who likes what and, why?” insights into modeling users’ personality based on image ‘likes”, *IEEE Transactions on Affective Computing* **9**(1), 130–143.  
**URL:** [https://doi.org/10.1109%2Ftaffc.2016.2581168](https://doi.org/10.1109/2Ftaffc.2016.2581168)
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. & Zhao, L. (2018), ‘Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey’, *Multi-media Tools and Applications* **78**(11), 15169–15211.  
**URL:** [https://doi.org/10.1007%2Fs11042-018-6894-4](https://doi.org/10.1007/2Fs11042-018-6894-4)
- Li, R. Z., Urbano, J. & Hanjalic, A. (2021), Leave no user behind: Towards improving the utility of recommender systems for non-mainstream users, *in* ‘Proceedings of the 14th ACM International Conference on Web Search and Data Mining’, ACM.  
**URL:** [https://doi.org/10.1145%2F3437963.3441769](https://doi.org/10.1145/2F3437963.3441769)

- Mandal, S. & Maiti, A. (2018), Explicit feedbacks meet with implicit feedbacks: A combined approach for recommendation system, *in* ‘Studies in Computational Intelligence’, Springer International Publishing, pp. 169–181.  
**URL:** <https://doi.org/10.1007%2F978-3-030-05414-4-14>
- McAllister, A., Naydenova, I. & Duc, Q. N. (2019), ‘Building a lda-based book recommender system’.  
**URL:** [https://humboldt-wi.github.io/blog/research/information\\_systems\\_1819/is\\_lda\\_final/#jaccard](https://humboldt-wi.github.io/blog/research/information_systems_1819/is_lda_final/#jaccard)
- Nagahi, M., Jaradat, R., Amrani, S. E., Hamilton, M. & Goerger, S. R. (2020), ‘Holistic and reductionist thinker: a comparison study based on individuals' skillset and personality types’, *International Journal of System of Systems Engineering* **10**(4), 337.  
**URL:** <https://doi.org/10.1504%2Fijsse.2020.112312>
- Paiva, F. A. P., Costa, J. A. F. & Silva, C. R. M. (2017), A personality-based recommender system for semantic searches in vehicles sales portals, *in* ‘Lecture Notes in Computer Science’, Springer International Publishing, pp. 600–612.  
**URL:** [https://doi.org/10.1007%2F978-3-319-59650-1\\_51](https://doi.org/10.1007%2F978-3-319-59650-1_51)
- Park, J.-H. (2017), ‘Resource recommender system based on psychological user type indicator’, *Journal of Ambient Intelligence and Humanized Computing* **10**(1), 27–39.  
**URL:** <https://doi.org/10.1007%2Fs12652-017-0583-4>
- Piedboeuf, F., Langlais, P. & Bourg, L. (2019), Personality extraction through LinkedIn, *in* ‘Advances in Artificial Intelligence’, Springer International Publishing, pp. 55–67.  
**URL:** [https://doi.org/10.1007%2F978-3-030-18305-9\\_5](https://doi.org/10.1007%2F978-3-030-18305-9_5)
- Rajapaksha, M. & Silva, T. (2019), Semantic information retrieval based on topic modeling and community interests mining, *in* ‘2019 Moratuwa Engineering Research Conference (MERCon)’, IEEE.  
**URL:** <https://doi.org/10.1109%2Fmercon.2019.8818935>
- Szmydt, M. (2021), ‘Contextual personality-aware recommender system versus big data recommender system’, *Business Information Systems* pp. 163–173.  
**URL:** <https://doi.org/10.52825%2Fbis.v1i.38>
- Tkalčić, M., Ferwerda, B., Hauger, D. & Schedl, M. (2015), Personality correlates for digital concert program notes, *in* ‘Lecture Notes in Computer Science’, Springer International Publishing, pp. 364–369.  
**URL:** [https://doi.org/10.1007%2F978-3-319-20267-9\\_32](https://doi.org/10.1007%2F978-3-319-20267-9_32)
- Woods RA, H. P. (2022), *Myers Brigg*, Treasure Island (FL): StatPearls Publishing.
- Xiong, S., Wang, K., Ji, D. & Wang, B. (2018), ‘A short text sentiment-topic model for product reviews’, *Neurocomputing* **297**, 94–102.  
**URL:** <https://doi.org/10.1016%2Fj.neucom.2018.02.034>
- Xu, K., Qi, G., Huang, J., Wu, T. & Fu, X. (2018), ‘Detecting bursts in sentiment-aware topics from social media’, *Knowledge-Based Systems* **141**, 44–54.  
**URL:** <https://doi.org/10.1016%2Fj.knosys.2017.11.007>
- Yang, A. (2015), Inferring business similarity from topic modeling [ latent dirichlet allocation and jaccard similarity applied to yelp reviews ].