Sentiment Analysis of K-pop Songs

Author:

Rose Maria George

Student ID: 22251254

Supervisor:

DR. JOSEPH TIMONEY

Head of the Department:

DR. JOSEPH TIMONEY

A dissertation submitted in partial fulfilment of the requirements

for the degree of

MSc in Data Science and Analytics

2022-2023

in the

Department of Computer Science,

Maynooth University

August 7, 2023

ABSTRACT

MSc in Data Science and Analytics

Title: Sentiment Analysis of K-pop Songs

by
ROSE MARIA GEORGE

This thesis explores the emotional landscape contained in the lyrics of K-pop songs through a Sentiment Analysis. The research uses a multi-stage approach, beginning with data preprocessing to ensure accurate analysis by removing stop words and explicit words. The VADER sentiment tool is used for sentiment analysis, and word clouds that depict positive and negative words help to better understand how sentiment is distributed in K-pop songs. By locating underlying themes in the song lyrics, Topic Modeling with Latent Dirichlet Allocation (LDA) enhances the analysis even further. Term Frequency-Inverse Document Frequency (TF-IDF) features are used in predictive modeling with machine learning algorithms like Logistic Regression, SVM, Naive Bayes, and the unsupervised method, BERT. Comparing their performances has important ramifications for predicting sentiment in K-pop songs. Additionally, the relationship between danceability and sentiment patterns is examined by comparing danceability scores and sentiment analysis. A deeper understanding of the emotional impact of music is also made possible by comparing K-pop with pop songs because it sheds light on the distinctive emotional expressions of these music genres and their influence on listeners. In addition to its contagious beats and mesmerizing performances, K-pop's magnetic allure lies in its capacity to evoke a wide range of emotions in listeners, creating a shared emotional experience that unites fans across the globe.

Keywords: K-pop, Sentiment Analysis, VADER, LDA, Logistic Regression, SVM, Naive Bayes, BERT, Danceability, Pop Music, Cultural Expressions.

Acknowledgment

I would like to express my sincere gratitude for the chance to work with Dr. Joseph Timoney, who is also my primary supervisor, as well as for his understanding and guidance throughout the project. Without his insightful suggestions and insightful criticism, this project would not have been possible. I want to thank the computer science department's technical and support staff at Maynooth University for their help. I also want to thank my supervisors for their assistance in getting my project finished.

Declaration

I hereby declare that the material I am submitting for assessment as a requirement for the MSc in Data Science and Analytics qualification program is entirely original and has not been modified in any way, with the exception of instances where it has been cited and acknowledged within the text of my work.

Signed: ROSE MARIA GEORGE Date: 07-08-2023

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Sentiment Analysis of Song Lyrics

In the field of Natural Language Processing (NLP), sentiment analysis is essential for comprehending and analyzing human emotions expressed in text. Humans are able to reason, deliberate, and articulate their opinions on a variety of topics, so it stands to reason that their emotions will be greatly influenced by the environment and will either be positive or negative. Sentiment analysis, a potent NLP tool, delves into understanding these subtle emotional nuances in textual data.

In the context of sentiment analysis of K-pop songs, this study aims to explore the emotional landscape in song lyrics. The results of this study have the potential to have significant ramifications for cross-cultural communication, marketing tactics, and the music business. In addition, examining the emotions expressed in K-pop songs can help us better understand the tastes and sentiments of a diverse global audience. This will help us comprehend the musical genre's cross-cultural influence and allure. This study aims to shed light on the profound impact that music has on influencing people's emotions and perceptions by analyzing the emotional content of K-pop lyrics, strengthening the bond between musicians and their devoted fan bases around the world.

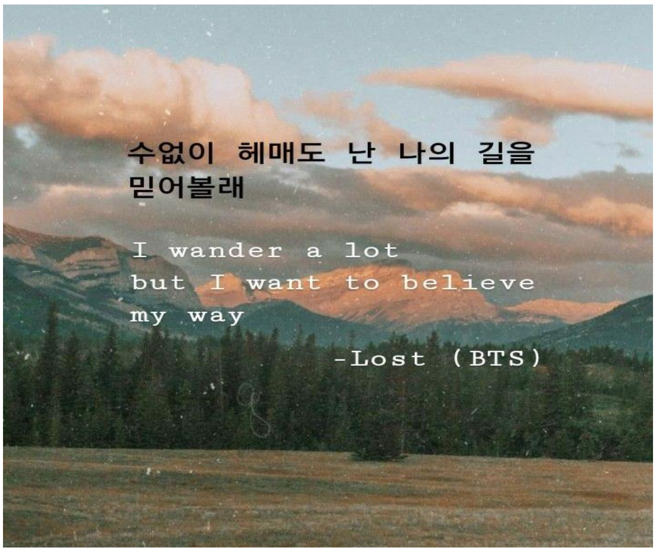

Figure 1. Emotional Landscape of K-pop Song Lyrics. Source (Pinterest)

1.2 Korean Language and its Significance in K-pop Sentiment Analysis

The Korean language, also known as "Hangul" in South Korea and "Chosn'gl" in North Korea, is an important component of the Korean Peninsula's cultural and linguistic landscape. Millions of people speak it in Korean diaspora communities all over the world, and it is the official language of both North and South Korea.

It was during the Joseon Dynasty in the 15th century that the distinctive writing system known as Hangul was developed. It was introduced by King Sejong the Great and a group of academics with the intention of raising literacy levels among Koreans. The development of Hangul was a ground-breaking linguistic accomplishment, and it is regarded as one of the most rational and effective writing systems in existence.

The grammar of the Korean language is renowned for its beauty and complexity, and it includes honorifics and different politeness levels that signify speakers' relationships and social status. It is an agglutinative language, which enables complex sentence structures and intricate expressions. Words are formed by adding suffixes to a base word.

Figure 2. K POP. Source (Wikipedia)

Korean is notable for its contrastive use of consonants and vowels, which are arranged into syllable blocks, in terms of phonetics. One initial consonant, one vowel, and possibly a final consonant makes up each block of syllables. Because of its distinctive phonological characteristics, Korean has a melodic and rhythmic quality that is ideal for the lyrical expressions found in K-pop songs.

The Korean language has become more widely known to international fans as K-pop has grown in popularity. In order to reach a wider audience, K-pop artists frequently incorporate English words and phrases into their songs. This increases the genre's international appeal and promotes intercultural communication. In order to accurately interpret the emotional themes and sentiments expressed through the lyrics of K-pop songs, it is crucial to understand the nuances of the Korean language.

1.3 The Evolution of K-Pop Culture

Korean pop, or K-pop, has become a global cultural phenomenon with a sizable fan base all over the world. K-pop, which has its roots in South Korea, is distinguished by its earworm melodies, gorgeous music videos, and expertly choreographed dance moves, making for an engrossing and immersive entertainment experience.

Early 1990s Western pop and hip-hop music, as well as traditional Korean musical elements, had a significant influence on the development of K-pop, which can be traced back to those times. But it wasn't until the late 1990s and the beginning of the 2000s that K-pop really started to take off, thanks to the appearance of well-known boy bands and girl groups like H.O.T, Sechs Kies, S.E.S, and Fin.K.L. These pioneering performers significantly contributed to the domestic and regional Asian success of K-pop

The establishment of the "Big 3" entertainment companies in the early 2000s—SM Entertainment, YG Entertainment, and JYP Entertainment—further fueled the development of the K-pop sector. With significant investments in production and training to produce polished and marketable artists, these agencies rose to prominence in the development and management of gifted idols. With the development of social media and online streaming services, K-pop crossed international borders and became popular in other countries. The worldwide success of PSY's "Gangnam Style" in 2012 elevated K-pop to the forefront of popular music worldwide, becoming the first YouTube video to surpass one billion views. Since then, K-pop has become a major force in world culture thanks to the enormous success of groups like BTS, EXO, BLACKPINK, and TWICE abroad.

Aspiring idols spend years honing their singing, dancing, and other abilities in the competitive K-pop industry before making their debut. Through this training, the artists become well-rounded performers with unique skills and personalities.

The success of K-pop can be attributed to its capacity for self-reinvention, which it has demonstrated by embracing a wide range of musical genres and styles as well as pushing the boundaries of creativity and innovation. Audiences of all ages and cultural backgrounds have been enthralled by the industry's focus on high-quality production, visually appealing concepts, and compelling storytelling. Another important factor in the success of K-pop is the fandom culture that surrounds it. Known as "K-pop stans," fans actively support their favorite artists on social media, through streaming, and by casting their votes in music award ceremonies.

Overall, the global impact of the K-pop music scene is expanding, breaking down barriers and promoting cross-cultural dialogue. With its distinctive fusion of music, visuals, and performance, K-pop has made a significant cultural impact on the world of entertainment.

1.4 Motivation

The emotional bond between performers and fans is strong and captivating in the exciting world of K-pop, a style that has won over hearts all over the world. K-pop music is a remarkable cultural phenomenon because its appeal cuts across linguistic and cultural divides. Only a few studies have concentrated on the sentiment analysis of K-pop songs, particularly in their native Korean language, despite the fact that there have been many studies analysing sentiment in songs from different music industries. Thus, this thesis sets out on a fascinating journey to investigate the emotional terrain found in K-pop lyrics.

The main goal of this study is to thoroughly analyse the sentiment of K-pop songs to identify the range of emotions that are conveyed in the lyrics. The study is divided into several phases, the first of which is the painstaking pre-processing of the data to ensure accurate and insightful analysis. Exploratory data analysis teaches important things about word frequency, which leads to a better understanding of the key themes and emotions present in K-pop songs. The VADER sentiment tool is used to conduct sentiment analysis, which enables data labelling and the creation of word clouds to visualize positive and negative words. In-depth analysis of the sentiment distribution in K-pop songs also sheds light on the overall emotional atmosphere of this dynamic musical genre.

Utilizing Term Frequency-Inverse Document Frequency (TF-IDF) features, a number of machine learning algorithms, such as Logistic Regression, Support Vector Machine (SVM), Naive Bayes and the unsupervised BERT model, are applied in the pursuit of predictive modeling. are applied in the pursuit of predictive modeling. In order to determine the most successful method for sentiment prediction in K-pop songs, the performance of these models is carefully compared. The thesis delves into statistical analysis, looking at patterns and trends in the emotional expressions found in K-pop lyrics. Additionally, the analysis gains depth and context from topic modeling using Latent Dirichlet Allocation (LDA), which identifies underlying themes and topics in the enormous corpus of K-pop song lyrics.

This research offers significant implications for comprehending the profound influence of K-pop songs on their global audience by revealing the emotional core of K-pop lyrics. The findings of this study contribute to a deeper understanding of cultural expressions that cross geographical boundaries in addition to advancing the field of sentiment analysis. Additionally, the conclusions drawn from this analysis have implications for the music industry, intercultural communication, and the enduring relationship between K-pop artists and their devoted fans around the world.

An enthralling journey into the world of emotions woven into K-pop songs unfolds through this thesis, revealing the vibrant tapestry that knits together artists and fans in a symphony of sentiments and melodies.

Chapter 2

Background Literature Review

Sentiment analysis, also known as opinion mining, is a significant area of research that aims to understand sentiments, emotions, and viewpoints from textual data. Sentiment analysis has gained popularity across a wide range of industries, including music analysis. This review of the literature focuses on the novel contributions of the current thesis that go beyond the purview of the reviewed literature while examining pertinent research papers on sentiment analysis in K-pop songs. This study seeks to advance our understanding of sentiment analysis as it pertains to K-pop songs by fusing existing knowledge and novel insights, offering significant implications for both the sentiment analysis field and the alluring world of K-pop music.

"KOSAC: A Full-fledged Korean Sentiment Analysis Corpus," a seminal work by Hayeon Jang, Munhyong Kim, and Hyopil Shin, makes an important contribution to the field of sentiment analysis, particularly when it comes to K-pop songs. The researchers present KOSAC, a comprehensive and meticulously curated Korean Sentiment Analysis Corpus. This corpus, which includes K-pop song lyrics, is an invaluable resource for sentiment analysis in the Korean language. KOSAC was inspired by the recognition of the distinct linguistic characteristics and emotional expressions found in K-pop music. It provides a rich and diverse collection of K-pop song lyrics sourced from various artists and groups spanning various genres and eras.

With this large dataset, it is possible to create and test sentiment analysis models that are specifically tailored to the complexities of K-pop song sentiments. Experts annotate the corpus extensively, ensuring accurate sentiment labelling and allowing for robust analysis and reliable benchmarking of sentiment analysis techniques. Furthermore, the availability of KOSAC to the research community encourages collaboration and advancements in the field by encouraging the exploration of novel methodologies and the development of innovative sentiment analysis tools for K-pop songs. With the help of this contribution, it is now possible to gain a deeper understanding of how K-pop affects audiences all over the world emotionally. The accessibility of KOSAC allows for a thorough investigation of the complex emotions conveyed in K-pop song lyrics, advancing knowledge of the emotional ties that unite artists and their devoted fanbase within the vibrant K-pop industry. [1].

The research by Humberto Corona and Michael P. O'Mahony, "An Exploration of Mood Classification in the Million Songs Dataset," is a thorough investigation of mood classification using a large music dataset. While the study does not specifically target K-pop songs, its findings have the potential to provide valuable insights into the emotional expressions inherent in this genre and can serve as a catalyst for mood-based sentiment analysis in the context of K-pop. The Million Songs Dataset is a rich source of musical information that makes it possible to investigate different musical qualities that are connected to mood. By examining elements like tempo, key, and harmonic content, Corona and O'Mahony are able to pinpoint patterns in music that correspond to different emotional states. These results may not be specific to K-pop songs, but they can offer a fundamental understanding of how mood is represented in song compositions.

Although these findings are general to music in general, they can provide a foundational understanding of mood representation in song compositions, which may be applicable to K-pop songs. Furthermore, the methodology and approaches used in the study for mood classification can inspire sentiment analysis can be applied to K-pop song lyrics. While sentiment analysis seeks to extract emotions and opinions from textual data, mood classification in music can provide a complementary perspective on the emotional nuances found in K-pop songs. Sentiment analysis models for K-pop songs may be able to incorporate mood-related features and concepts particular to K-pop music by drawing on Corona and O'Mahony's research findings. The accuracy and depth of sentiment analysis in K-pop songs may be enhanced by this integration [2].

K-pop songs from the past fifty years are examined in "Quantitative Analysis of a Half-Century of K-Pop Songs: Association Rule Analysis of Lyrics and Social Network Analysis of Singers and Composers," a study by Yeawon Yoo and Yonghan Ju. While their research does not primarily focus on sentiment analysis, it does provide valuable insights into the lyrical patterns and social relationships prevalent in K-pop songs, which can significantly influence the conveyed sentiments within the genre. To find meaningful connections between the words and phrases used in K-pop songs, Yeawon Yoo and Yonghan Ju use association rule analysis of the lyrics. By highlighting recurring themes, feelings, and sentiments in the lyrics, this analysis may reflect the songs predominate emotional expressions and help the listeners understand the songs' overall message.

Furthermore, the social network analysis of singers and composers reveals the collaborative dynamics within the K-pop industry. The study emphasizes the interconnectedness and influence of artists and creators on each other's work, demonstrating how collective emotions and experiences can be woven into the fabric of K-pop songs. These social relationships have the potential to influence the emotions expressed in the lyrics and performances, thereby shaping the sentiment perceived by the audience. Yeawon Yoo and Yonghan Ju's research lays the groundwork for understanding the emotional undercurrents in K-pop songs, even though sentiment analysis is not the main focus. In order to better capture the complex emotional expressions conveyed through music, sentiment analysis models can make use of the identified lyrical patterns and social relationships as contextual cues [3].

“Bidirectional Encoder Representations from Transformers (BERT): A Sentiment Analysis Odyssey” by Shivaji Alaparthi and Manit Mishra presents a thorough investigation of BERT's use in sentiment analysis. This study explores the ground-breaking features of BERT's bidirectional context representation, which makes it possible to capture intricate contextual relationships within sentences. BERT exhibits remarkable performance, outperforming conventional models and other deep learning architectures in sentiment classification by pre-training on large corpora and optimizing on sentiment analysis tasks. The study discusses the efficiency of BERT in a variety of scenarios, including various languages and domains, and it displays the methodologies and techniques used to adapt BERT for sentiment analysis. When applying sentiment analysis to K-pop songs, the research's insights can be very helpful.

The accuracy and comprehension of the emotional and sentimental content expressed by K-pop artists in their songs may be improved by BERT's capacity to grasp intricate linguistic structures and contextual nuance. Choosing the best model for a particular sentiment analysis of K-pop song lyrics is made easier by the paper's comparison of BERT and other sentiment analysis techniques. Overall, the study by Shivaji Alaparthi and Manit Mishra is a useful tool for understanding BERT's development in sentiment analysis. Their discoveries advance sentiment analysis methodologies and may improve our comprehension of the emotions conveyed in K-pop song lyrics. This study emphasizes the value of transformer-based models like BERT in the area of sentiment analysis and natural language understanding, laying a strong foundation for future research and advancements in the sentiment analysis of K-pop songs [4].

Aside from an extensive review of the literature, the current thesis makes several original and innovative contributions to the field of sentiment analysis of K-pop songs. To begin, the thesis introduces a novel approach for sentiment analysis in K-pop songs by combining VADER sentiment analysis with BERT-based unsupervised training. This combination of two powerful techniques enables a more comprehensive and nuanced analysis of the emotional nuances present in K-pop song lyrics. The method significantly improves sentiment classification accuracy and precision by leveraging the strengths of VADER's rule-based approach and BERT's deep contextual understanding. This novel approach allows to delve deeper into the complex emotions conveyed through the enchanting verses of K-pop songs, providing valuable insights into the artists' intentions and the sentiments that resonate with their global fanbase.

The thesis also investigates the relationship between sentiments and danceability scores in K-pop songs. This fascinating investigation sheds light on the interplay between emotions and musical characteristics, revealing how sentiments can influence the danceability of K-pop songs. Understanding this link can have significant implications for music producers and artists, as it can help them create songs that elicit specific emotions and engage listeners on a deeper level.

Furthermore, the thesis compares various sentiment analysis models, such as logistic regression, SVM, Naive Bayes, and VADER-BERT. This rigorous evaluation aids in the selection of the most effective approach for sentiment classification in K-pop songs, taking accuracy and performance metrics into account. By identifying the best model, sentiment analysis can be streamlined to accurately capture and analyse the emotional nuances conveyed through K-pop lyrics.

Finally, this thesis not only provides a comprehensive review of existing research on sentiment analysis of K-pop songs, but it also significantly contributes to the field through novel and ground-breaking approaches. Through the enchanting verses of K-pop songs, the thesis sheds new light on the emotional connections between artists and their global fanbase by combining VADER and BERT for sentiment analysis, exploring the correlation between sentiments and danceability, and conducting a comprehensive model comparison. With the development of sentiment analysis in K-pop, music's profound emotional impact finds its voice, resonating with hearts and souls everywhere and fostering deeper emotional connections through the universal language of music. This study broadens the field of research, deepens understanding of the emotional content of K-pop songs, and strengthens relationships between artists and fans. It reveals K-pop's emotive influence on the global stage and paves the way for future sentiment analysis advancements with its originality and insightfulness.

Chapter 3

Sentiment Analysis Techniques

3.1 Overview

The goal of sentiment analysis, a task in natural language processing (NLP), is to ascertain the sentiment or emotional tone expressed in a text. The analysis of customer feedback, market research, and social media monitoring are just a few of the many industries that use it extensively. Sentiment analysis is used to comprehend the emotional content that K-Pop songs convey to their listeners and how that content connects with them. Sentiment analysis is a field that uses a variety of methods.

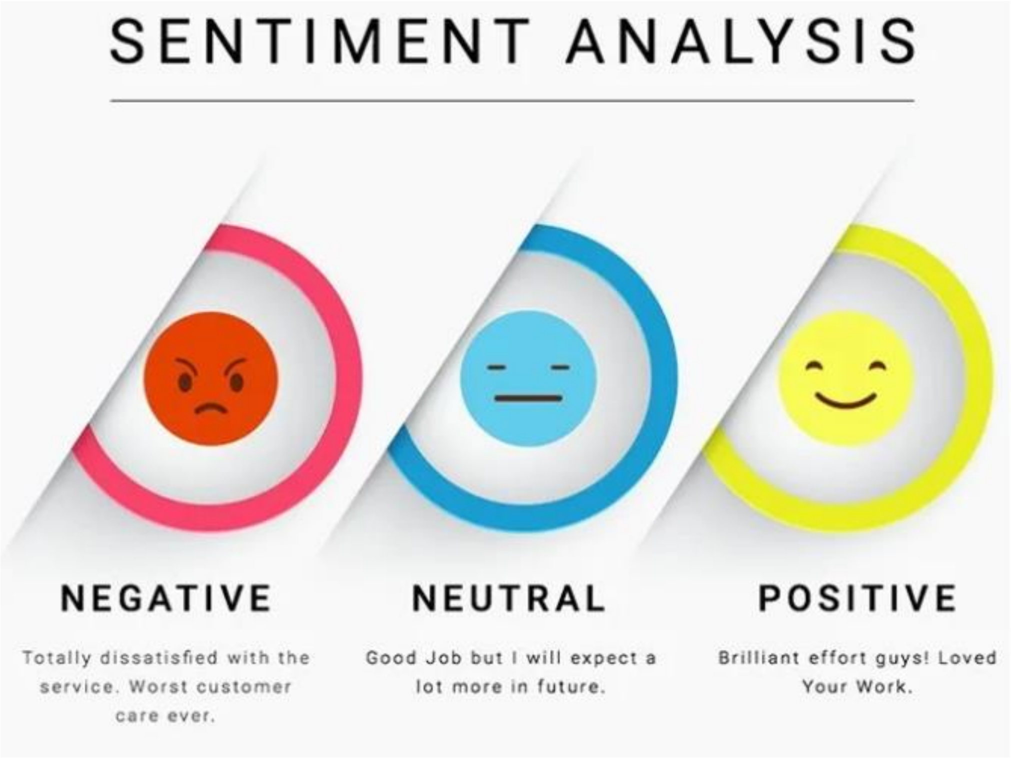
SENTIMENT ANALYSIS
NEGATIVE
Totally dissatisfied with the service. Worst customer care ever.
NEUTRAL
Good Job but I will expect a lot more in future.
POSITIVE
Brilliant effort guys! Loved Your Work.

Figure 3. Sentiment Analysis. Source (Wikipedia)

3.2 Types of Sentiment Analysis Approaches

There are various kinds of sentiment analysis techniques, and each one uses a different methodology to decipher and classify the sentiments in text. The methods that are most frequently employed are:

3.2.1 Lexicon-based approaches: Lexicon-based sentiment analysis uses lexicons or sentiment dictionaries that have already been created and contain words along with the corresponding sentiment scores. Each word in the text is given a polarity score by the analyzer, and the sum or average of these scores is used to calculate the text's overall sentiment. Lexicon-based methods are effective and don't need a lot of labelled training data. A well-known lexicon-based tool that works well with short, informal expressions and social media text is VADER Sentiment Analyzer.

Lexicon-based approaches make use of dictionaries or lexicons of words with associated sentiment scores (such as positive, negative, or neutral). The overall sentiment of a text is determined using the sentiment scores of the individual words, and each word is given a polarity score. These methods can be helpful for quick sentiment analysis and are reasonably simple to put into practice. They might not, however, fully express the context or subtleties of the feelings expressed in a text. The VADER Sentiment Analyzer was used in this section to analyze the sentiment in K-pop songs.

3.2.1.1 VADER Sentiment Analyzer

Specifically created for social media text, the VADER (Valence Aware Dictionary and Sentiment Reasoner) Sentiment Analyzer is a lexicon and rule-based tool. It can evaluate sentiment in a sentence or document and assign sentiment scores to words based on their valence and intensity.

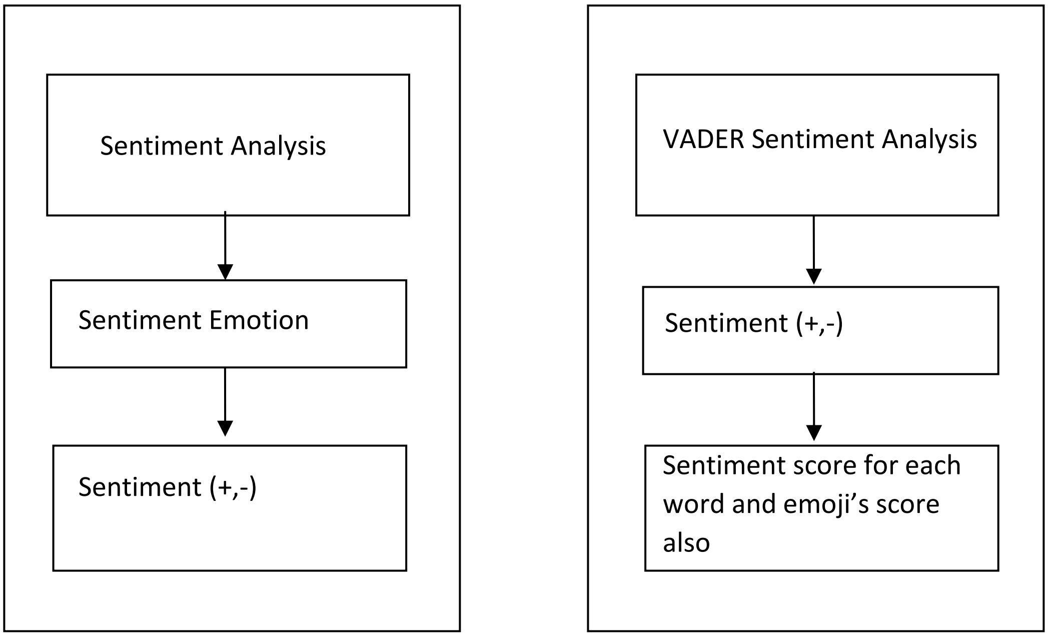

Table 1. Sentiment Analysis vs VADER Sentiment Analysis

3.2.1.2 Interpreting VADER Sentiment Scores

The scoring methodology and its relationship to sentiment must be understood in order to interpret the VADER sentiment scores. Positive values indicate positive sentiment, negative values indicate negative sentiment, and values close to zero represent neutral sentiment. The VADER Analyzer assigns sentiment scores in a continuous range from -1 to +1. The score's absolute magnitude, which is larger for stronger absolute values, indicates how intense the sentiment is.

The VADER Analyzer aggregates the polarity scores of all the words in the lyrics, taking into account their intensities and the context in which they appear, to determine the overall sentiment of a song. The song is deemed to have a positive sentiment if the sum of the polarity scores is primarily positive. On the other hand, the song is thought to have a negative sentiment if the sum is predominately negative. The song is deemed to have a neutral sentiment when the sum is close to zero.

3.2.1.3 Strengths and Limitations of VADER Sentiment Analysis

In the context of sentiment analysis for K-pop songs, the VADER Sentiment Analyzer demonstrated a number of advantages. Its simplicity and usability, which make it accessible to researchers and practitioners without a thorough understanding of machine learning techniques, is one of its main advantages. Furthermore, because K-pop song lyrics frequently contain emoticons, slang, and informal language, VADER was created specifically to analyze social media text.

The VADER Sentiment Analyzer's capacity to detect subtle nuances in sentiment in text is another strength. It allows for a more nuanced understanding of the overall sentiment of a song by taking into account the intensity of sentiment expressed by each word. Additionally, VADER offers distinct scores for sentiments that are positive, negative, and neutral, allowing for a more detailed analysis of the lyrics.

The VADER Sentiment Analyzer has some restrictions despite its advantages. Its reliance on predefined lexicons represents one significant limitation. Despite being comprehensive, the lexicon might not include all of the precise words or expressions used in K-pop songs. As a result, VADER might not correctly interpret some slang, recently created terms, or words with multiple meanings, which could result in sentiment being misclassified.

Another difficulty with using VADER is how poorly it handles irony and sarcasm. VADER may have trouble accurately interpreting feelings conveyed through sarcasm or other figurative language because it only performs word-level analysis.

3.2.2 Machine learning models

Using labelled data, supervised learning models are trained for sentiment analysis using machine learning. During training, these models discover patterns and connections between textual elements and corresponding sentiments. For classifying sentiment, common machine learning algorithms include logistic regression, support vector machines (SVM), and Naive Bayes. Machine learning models are extremely adaptable for various types of text analysis tasks and can handle complex patterns in text data.

The development of predictive models that can categorize text into positive, negative, or neutral sentiments based on patterns in the data is made possible by machine learning, which is essential to sentiment analysis. The fundamentals of machine learning, the significance of using it for sentiment analysis, and the specific machine learning methods used in the study are all covered in this section. The two main learning paradigms used in machine learning are supervised learning and unsupervised learning.

3.2.2.1 Supervised Learning:

In supervised learning, the algorithm is trained using labeled data, which means that the input data is paired with corresponding output labels. The objective of supervised learning is to develop a relationship between the input features and the target labels that will enable the algorithm to predict outcomes for brand-new, untainted data. supervised learning algorithms are frequently used in sentiment analysis to train models that categorize text into predefined sentiment categories (such as positive, negative, or neutral) based on labeled examples.

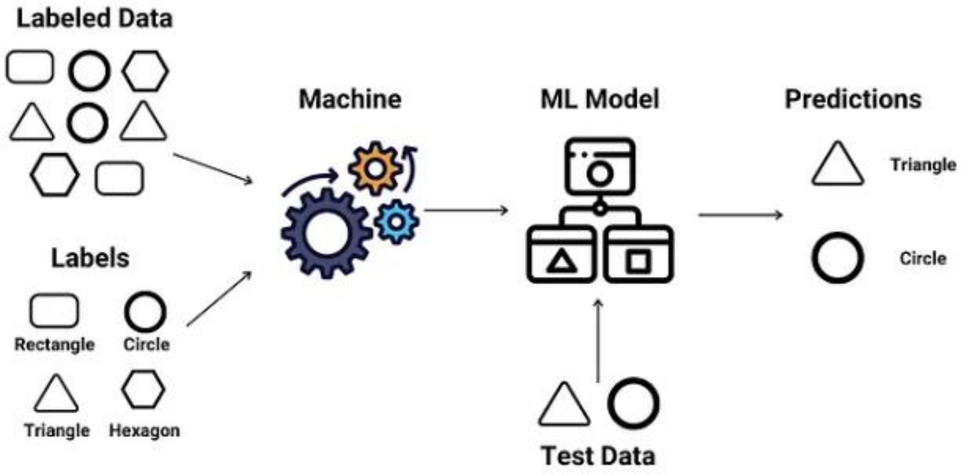

Figure 4. Supervised Learning

i. Logistic Regression:

It is possible to classify the sentiment in K-pop songs as either positive or negative using the widely used supervised learning algorithm, logistic regression, which handles binary classification tasks. A sigmoid function is used by the logistic regression model to determine how the binary sentiment labels and the input features (textual data from the lyrics) are related. This function converts the continuous output to a probability value, which is then thresholder to identify whether K-pop songs have a positive or negative message. A practical option for sentiment analysis of K-pop lyrics is logistic regression because it is computationally effective and understandable.

ii. Support Vector Machines (SVM):

Another well-liked classification algorithm, Support Vector Machine (SVM), is capable of performing binary and multi-class classification tasks. In K-pop song sentiment analysis, SVM learns to define a decision boundary that clearly distinguishes instances of positive and negative sentiment in the feature space. New K-pop songs will have better generalization and improved sentiment prediction accuracy thanks to SVM, which aims to maximize the margin between the two classes. Relationships with non-linearity can be handled by SVM

iii. Naive Bayes:

Based on Bayes' theorem, the Naive Bayes classifier excels at tasks requiring text classification, such as sentiment analysis. Given the class label, it operates under the presumption that the features are independent of conditions. Naive Bayes still performs remarkably well in the sentiment prediction for K-pop songs, despite the fact that real-world data frequently contradicts this assumption. It is effective in situations with little labeled data because it is computationally efficient, requires little data preprocessing, and uses few resources. Naive Bayes is a useful addition to sentiment analysis tools for K-pop music because it can classify K-pop songs into positive or negative sentiments depending on the likelihood that a word will occur in each class.

3.2.2.2 Unsupervised Learning:

Contrarily, unsupervised learning involves training the algorithm on unlabeled data without any clear output labels. Unsupervised learning seeks out patterns or structure in the data, such as clustering related data points or lowering the dimensionality of the data. Although there are many uses for unsupervised learning in data exploration and feature extraction, sentiment analysis is a field where labeled data is frequently available for training supervised models, so its use is less common.

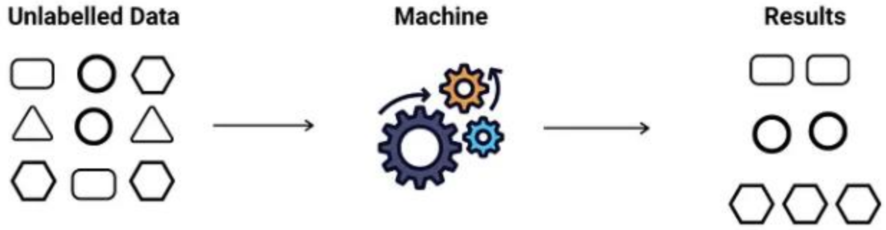

Figure 5. Unsupervised Learning

i. BERT Model:

Modern deep learning techniques, such as Bidirectional Encoder Representations from Transformers (BERT), have completely changed sentiment analysis and other natural language processing tasks. BERT is a contextualized word embedding technique that, in contrast to conventional models like Naive Bayes, takes into account the surrounding words to understand the meaning of a word in a given context. BERT can capture intricate emotional nuances and dependencies within the lyrical narrative of K-pop songs thanks to contextualization, which results in more precise sentiment predictions.

The BERT-based transformer model has demonstrated to be highly efficient in sentiment analysis for K-pop songs, as it can accurately identify complex emotional patterns within the lyrics. BERT outperforms conventional methods even when real-world data conflicts with the independence assumption frequently made by Naive Bayes. This is because BERT can use large-scale pretraining data and adapt to particular tasks through fine-tuning.

Additionally, the BERT method has a number of benefits for sentiment analysis in K-pop music. First of all, it uses little data preprocessing, saving time and money. Furthermore, because it is unsupervised, it can handle situations with little labeled data, which is a common situation in the analysis of K-pop songs where large labeled datasets might not always be accessible.

The accuracy and precision of sentiment classification have been greatly improved by incorporating the BERT approach into sentiment analysis tools for K-pop songs. BERT has developed into a useful tool for researchers and enthusiasts alike, providing a more thorough understanding of the sentiments prevalent in this captivating genre by comprehending the intricate emotional connections conveyed through the enchanting verses of K-pop songs.

3.2.3 Tools Used

i. Python as a programming language

Python was created by Guido van Rossum as a general-purpose programming language in the late 1980s. It is widely used in data science and engineering and has a sizable and growing user base. Python can be written as procedural, functional, and object-oriented code.

Python has a number of advantages over other programming languages, including the following:

Simple – Comparatively speaking, Python is a programming language that is comparatively very simple and user-friendly.

Compatibility – Python's compatibility with such a broad range of hardware and operating systems allows it to be used successfully on a variety of them at the same time.

Open Source – Python is a programming language that is open-source. This suggests that its open-source nature and substantial user base have accelerated the language's evolution.

Object Oriented – Python's object-oriented design makes it easier for programmers to build server-side and web-based applications.

Support for multiple libraries – Numerous libraries devoted to data science and machine learning have been developed over time using the Python programming language. TensorFlow, Pandas, NumPy, SciPy, NumPy, and Scikit-Learn are a few of the most well-known libraries. The libraries are simple to use and simple to incorporate into already developed applications.

With the Python package management tool, also known as PIP, packages can be installed quickly and easily. Additionally, Python enables us to build distinct virtual environments for every application, which makes it simpler for us to maintain distinct packages.

ii. Python Libraries

In this project, various Python libraries are being used. Others need to be manually installed using the Python PIP package manager, although some of these libraries can be installed automatically using the Python interpreter.

Pandas: The Pandas library is essential for data scientists. A machine learning library that provides a range of analysis tools and flexible high-level data structures. The program makes it easier to manipulate, purge, and analyze data. Data conversion, visualization, aggregate operations, iteration, concatenation, sorting, and re-indexing are all supported by Pandas.

NumPy: Numpy is an abbreviation for "Numerical Python." It is the one that is most frequently used. This well-liked machine learning library supports large matrices and multi-dimensional data. For quick calculations, it has built-in mathematical functions. For many internal tensor operations, libraries like TensorFlow also use NumPy. The Array Interface is one of this library's key features.

OS: The OS module in Python offers resources for interacting with the operating system. OS is one of the common utility modules for Python. A portable way to access operating system-specific functionality is provided by this module. There are many ways to interact with the file system thanks to the os and os path modules.

TensorFlow: An open-source software library is TensorFlow. In order to conduct machine learning and deep neural network research, Google's Brain Team created TensorFlow as a component of its Machine Intelligence research organization. However, the system is sufficiently general to be useful in other domains as well.

Keras: The utility library for NumPy has functions for manipulating arrays. A NumPy array (or) vector of integers representing different categories can be converted into a NumPy array (or) matrix with binary values and a number of columns equal to the number of categories by using the method to categorical ().

Scikit-learn: An integrated interface for machine learning, pre-processing, cross-validation, and visualization algorithms is offered by the open-source Python library known as Scikit-learn.

The features of Scikit-Salient Learn are as follows:

Effective and simple to use data mining and analysis tools. Support vector machines, random forests, gradient boosting, k-means, and other algorithms for classification, regression, and clustering are some examples of these algorithms. readily available and adaptable to different contexts.

- Powered by SciPy, NumPy, and Matplotlib

- Open source, usable for business, under the BSD license.

Plotly: The Python Plotly Library is an open-source library that can be used to quickly and simply visualize and understand data. The various plot types that Plotly supports include line charts, scatter plots, histograms, and cox plots.

Plotly's hover tool features can be used to find any outliers or anomalies in a large number of data points.

- The ability to fully customize graphs increases their significance and readability for others and appeals to a wide range of audiences.

MinMaxScaler: If there are negative values in the dataset, MinMaxScaler scales all data features in the [-1, 1] range. All inliers in the constrained range [0, 0.005] are compressed by this scaling. Due to the impact of the outliers during the computation of the empirical mean and standard deviation, StandardScaler does not ensure balanced feature scales in the presence of outliers. The range of feature values is consequently narrowed.

Matplotlib: The Matplotlib library for Python is excellent for 2D array plotting. A multi-platform data visualization library built on NumPy arrays, Matplotlib integrates with the larger SciPy stack. It was first delivered by John Hunter in 2002. Access to enormous amounts of data is made possible by visualization in understandable formats. In Matplotlib, you can plot lines, bars, scatterplots, histograms, and more.

Seaborn: Python's Seaborn visualization library can be used to plot statistical graphics. The default color schemes and styles make statistical plots more visually appealing. It is based on the Matplotlib library and has a close relationship with Pandas data structures. Visualization is emphasized by Seaborn as the key to understanding and exploring data. To help us better understand the dataset, it offers dataset-oriented APIs that let us switch between different visual representations for the same variables.

NLTK: For text processing, tokenization, stemming, part-of-speech tagging, and sentiment analysis, the Natural Language Toolkit offers a number of tools.

These Python libraries played a key role in the data processing and analysis, feature extraction, machine learning model construction, and sentiment analysis of K-pop song lyrics. The study made use of these libraries' advantages to gain insightful understandings into the feelings and attitudes conveyed in the songs, adding to a thorough sentiment analysis of the K-pop music genre.

Data was effectively organized and prepared for analysis using Pandas. During model training, NumPy made it simple to perform mathematical operations on large arrays. Data management and file handling became easier to access thanks to Python's OS module, which enabled seamless interaction with the operating system.

The NLTK provided crucial text processing, tokenization, and sentiment analysis tools for natural language processing. The sentiment intensity of K-pop song lyrics was evaluated using the SentimentIntensityAnalyzer from NLTK's VADER module.

TensorFlow and Keras were used in machine learning to create and train deep learning models for sentiment classification. For classification and regression tasks, Scikit-learn provided a wide variety of machine learning algorithms, including preprocessing methods and model evaluation.

Plotly, Matplotlib, and Seaborn were used to effectively visualize the outcomes and sentiment patterns. The interactive capabilities of Plotly made it simple to spot outliers and anomalies in the data, and Matplotlib and Seaborn offered a wide range of plot types and statistical graphics for thorough data visualization.

Finally, the combination of these potent Python libraries made it possible to successfully implement sentiment analysis on K-pop song lyrics, giving valuable insights into the emotions and sentiments conveyed in the songs.

Chapter 4

Evaluation Metrices for Sentiment Analysis Models

Evaluation metrics are essential tools in sentiment analysis used to judge the performance and effectiveness of machine learning models used for sentiment classification in K-pop songs. These metrics help researchers choose the most appropriate approach by giving them important information about the model's accuracy in sentiment prediction. Quantifying the performance of the model with evaluation metrics enables understanding its benefits and drawbacks.

5.1. Confusion Matrix

An essential assessment tool that provides a thorough picture of the model's performance is the confusion matrix. It lists the total number of accurate and inaccurate predictions the model made for each sentiment class. The True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) metrics make up the matrix. TP stands for the proportion of positive sentiments that the model correctly predicted, TN for the proportion of negative sentiments that it correctly predicted, FP for the proportion of negative sentiments that were incorrectly classified as positive, and FN for the proportion of positive sentiments that were incorrectly classified as negative. Researchers can calculate additional evaluation metrics, such as precision, recall, and F1-score, by analyzing the confusion matrix, giving them a more in-depth understanding of the model's predictive abilities.

The effectiveness of machine learning models for classification is mainly assessed using these three metrics. These values are computed using the following terms,

True positive (TP) = Number of positive cases correctly classified as True

False positive (FP) = Number of negative cases incorrectly classified as True

True negative (TN) = Number of negative cases correctly identified as False

False negative (FN) = Number of positive cases incorrectly identified as False Accuracy, Proportion of true positive and true negative = (TP+TN)/(TP+TN+FP+FN)

5.2 Accuracy

A key evaluation metric for gauging the overall effectiveness of sentiment analysis models is accuracy. It shows the proportion of sentiment labels—both positive and negative—that were correctly predicted to all the predictions the model made. While accuracy provides a broad sense of how well the model performs, it might not be adequate to judge the model's performance in specific situations, especially when working with datasets that are unbalanced. When one sentiment class significantly outweighs the other in such circumstances, accuracy might not be an accurate reflection of the model's performance. Researchers frequently take into account additional metrics like precision, recall, and AUC to obtain a more thorough evaluation.

5.3 Precision

When dealing with positive sentiment predictions, precision is a key evaluation metric in sentiment analysis. It assesses the consistency of the model's positive sentiment predictions across all instances that the model has classified as positive. In essence, precision provides an answer to the query of how many of the model's optimistic predictions were accurate. A higher precision score means the model is correctly identifying positive sentiments and avoiding misclassifying negative sentiments as positive, which means it is making fewer false positive errors. Precision offers helpful insights into the model's capacity to precisely identify positive emotions and minimizes the risk of mistakenly attributing positive sentiments to negative expressions in the context of sentiment analysis of K-pop songs, where the emotional nuances can be diverse.

5.4 Recall

Another crucial metric in sentiment analysis is recall, which is also known as sensitivity or true positive rate. It is especially important for evaluating positive sentiment identification. It evaluates how well the model can distinguish between positive emotions among all the real positive examples in the dataset. Recall provides a straightforward answer to the query of how well the model captures all the positive sentiments present in the data. A higher recall score shows that the model effectively recognizes positive sentiments and minimizes false negatives, where positive sentiments are misclassified as negative. Recall provides important insights into the model's capacity to fully capture positive sentiments, ensuring that no positive expressions are missed, which is crucial for sentiment analysis of K-pop songs.

5.5 F1 Score

Precision and recall are both considered when calculating the F1 score, making it a balanced metric. It balances false positives and false negatives by calculating the harmonic mean of precision and recall. When dealing with imbalanced datasets, where one sentiment class outweighs the other, the F1 score is especially helpful. The F1 score helps strike a balance between accurately identifying positive sentiments and minimizing both false positives and false negatives by taking into account both precision and recall. A higher F1 score shows that the model successfully manages positive and negative sentiment, striking a good balance between precision and recall. The F1 score aids in evaluating the overall performance of the model for sentiment analysis of K-pop songs, where emotions can be complex and diverse.

5.6 Area Under the Curve (AUC)

The model's effectiveness in binary sentiment classification tasks is evaluated using the Area Under the Curve (AUC), which is a useful evaluation metric. By plotting the Receiver Operating Characteristic (ROC) curve, AUC assesses the model's capacity to distinguish between positive and negative sentiments. The ROC curve demonstrates how the discrimination threshold of the model affects the trade-off between the true positive rate (recall) and the false positive rate. The AUC provides a single scalar value that quantifies the model's capacity to rank positive sentiments higher than negative ones. It represents the model's overall performance across various discrimination thresholds. With a value of 1, a perfect classifier, a higher AUC indicates better model performance.

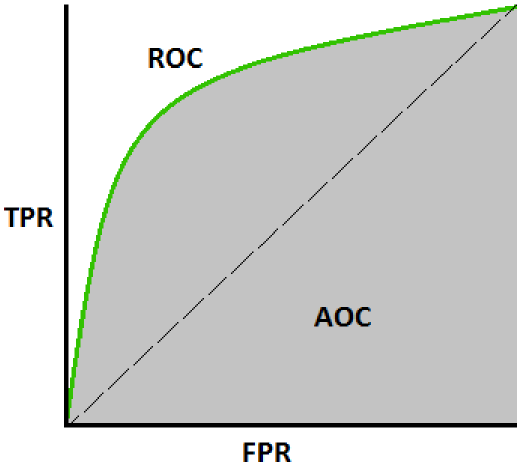

Figure 6. Area under the ROC Curve

In summary, evaluation metrics are crucial for assessing the effectiveness of sentiment analysis models and directing the choice of the most efficient strategy for sentiment classification in K-pop songs. While accuracy and AUC offer helpful insights into the model's overall performance and discrimination ability, the confusion matrix offers a detailed understanding of the model's predictive capabilities. Researchers can optimize sentiment analysis models and gain deeper insights into the emotional content of K-pop lyrics by utilizing these evaluation metrics and making well-informed decisions.

Chapter 5

Data Collection and Preprocessing

This chapter discusses the methods used to gather the data and create the dataset. Additionally, it covers both general data pre-processing and specific pre-processing for each model. In order to find any trends or patterns in the lyrics and in the sentiments that were included using human raters, the pre-processed data is analysed.

5.1 DATASET COLLECTION

There were several difficulties in obtaining a well-organized and structured dataset of Korean language songs for sentiment analysis. It was initially attempted to scrape song lyrics from different K-pop lyrics websites. These attempts, however, were hampered by API access problems and restrictions on how many songs could be fetched from particular platforms. The majority of K-pop song sentiment analysis studies translated the lyrics into English before analysis. So it was decided to manually build the dataset from scratch, using two primary sources:

5.1.2 Manual Collection from kpoplyrics.net Website

Despite difficulties encountered during web scraping, the goal of producing a high-quality dataset persisted. The dataset was primarily curated using manual collection, ensuring a thorough representation of the rich and varied K-pop music scene. The goal was to record a wide variety of K-pop songs, from classics from the previous generation, like Seo Taiji and Boys, to the wildly popular songs from the current generation, including songs from different K-pop idols, solo performers, boybands, and girlbands.

Each song's lyrics were painstakingly collected from the kpoplyrics.net website and saved in TXT format, with filenames that correctly reflected the names of the individual songs. The dataset aimed to include a variety of musical genres and themes, reflecting the ever-evolving nature of K-pop music [14].

Although the manual collection process required a lot of time and work, it was crucial to ensuring the accuracy and dependability of the dataset. The dataset established the framework for thorough sentiment analysis, enabling an in-depth exploration of the emotions and sentiments expressed within the genre. It did this by compiling a diverse collection of K-pop songs. The dataset's broad coverage of songs made it possible to gain understanding of how K-pop music has influenced global audiences and different cultures.

As a result, the dataset was a useful tool for academics and fans who were interested in the sentiments expressed in K-pop song lyrics. It contributed to a deeper understanding of the artistic and cultural significance of the genre by giving a comprehensive understanding of how sentiments and themes have changed from K-pop's beginnings to the present.

5.1.3. Integration of Data from Melon Streaming Platform

The Melon streaming platform's data was thoughtfully incorporated into the data collection process to further increase the dataset's richness and diversity. The prestigious Melon platform—a well-known music streaming service in South Korea known for its extensive music database—was accessed using a web scraping technique to obtain important data about the top K-pop songs for specific years.

A Python notebook created by a community member and made available on GitHub served as an excellent tool for this. The notebook was equipped with the program needed to effectively scrape websites using the Melon platform. Each song's lyrics were downloaded asynchronously from the Melon platform using the 'get_melon_lyrics' function. This method made it possible to retrieve lyrics data quickly without having to wait for one request to finish before sending out the next.

The dataset library was used to arrange and store the retrieved song information, including the lyrics, in a personal SQLite database. The Melon notebook was adjusted as necessary to cater the data collection procedure to particular needs [16].

The collaborative efforts and the Melon notebook's customizations led to the creation of a comprehensive and varied dataset that captures the dynamic world of K-pop music. With the help of this enriched dataset, a thorough sentiment analysis of K-pop song lyrics has been possible, providing insightful knowledge into the feelings, attitudes, and themes that underlie this popular music genre. The dataset was increased in terms of the number of songs and also included songs that were popular and trending during particular years by incorporating data from the Melon platform. As a result of this addition, the dataset gained a temporal dimension that allowed for the analysis of sentimental trends in K-pop music over time.

5.2 DATA PRE-PROCESSING

To make sure that the data is clear, organized, and appropriate for analysis, data preprocessing is a crucial step in any project involving data analysis or machine learning. In this project, a dataset of K-pop song lyrics was prepared for sentiment analysis using a variety of data preprocessing techniques. The text data had to be cleaned, stop words had to be removed, missing values had to be handled, and duplicates had to be removed.

5.2.1 Text Cleansing:

A custom function called text cleansing was put into place to clean up the text data. The text was converted to lowercase, any brackets were removed, new line breaks were replaced with spaces, punctuation was removed, extra whitespace was removed, and broken words were removed. By standardizing the text, these steps have made it simpler to compare and analyse the sentiments expressed in various songs.

5.2.2 Stop word Removal:

Stop words are frequently used words in a language that don't significantly add to the text's overall meaning. With the help of the remove stop words function, these words were eliminated. A customized list of explicit stop words was also incorporated into the procedure to account for terms that are specific to a given domain and may affect sentiment analysis results.

5.2.3 Handling Missing Values and Duplicates:

The dataset was examined for missing values to ensure data integrity, and any rows that contained missing data were removed. In order to prevent bias or redundancy in the analysis, duplicates were also eliminated from the dataset.

5.2.4 Data Encoding:

Encoding problems were resolved during the data collection process by experimenting with various encodings (utf-8, latin-1, and utf-16) until the data could be correctly read into the Data Frame. Having the appropriate encoding enabled easy data fusion and analysis.

5.2.5 Text Encoding Cleanup:

The 'lyrics' column underwent additional cleaning using regular expressions to address particular encoding artifacts found in the text data. The accuracy and coherence of the text were ensured by resolving the special character issues in this step.

5.2.6 Handling Bilingual Content:

The presence of bilingual content in the lyrics of K-pop songs presented one of the special challenges encountered during data preprocessing. English words or phrases are frequently incorporated into the beginning or end of Korean songs. The dataset's bilingual nature necessitated careful preprocessing to preserve the integrity of both languages.

5.2.7 Translation and Language Consistency:

The Google Translate API was used to obtain English translations of the Korean songs in order to maintain linguistic consistency and enable meaningful comparison and analysis. By standardizing the representation of the lyrics in the dataset, this step made it possible to conduct sentiment analysis using a single set of linguistic constructs. The process of translation made sure that the emotional expressions used in both languages were precisely recorded and matched for analysis.

5.2.8 Data Integrity and Validation:

Throughout the preprocessing phase, data validity and integrity were of utmost importance. The song lyrics, translations, and other pertinent information have all undergone extensive verification. The dataset's dependability was ensured by manual collection from the kpoplyrics.net website, which involved a meticulous procedure for compiling real K-pop song lyrics.

5.2.9 Addressing Technical Limitations:

Innovative solutions were needed because there were no databases of Korean song lyrics already in existence and web scraping was hampered by technical issues. To produce a complete dataset, a combination of manual collection and web scraping from the Melon streaming platform was used. Although manual collection took time, it guaranteed the accuracy and richness of the dataset. The dataset's scope was widened concurrently by web scraping, which offered an additional source of the best K-pop songs for particular years.

5.2.10 Data Quality Assurance:

Data quality control was essential during the preprocessing stage. This required careful handling of missing values, the eradication of duplicates, and the resolution of encoding problems. To ensure the accuracy of the sentiment analysis results, the cleaned and processed dataset underwent stringent quality checks to verify its integrity and consistency.

5.2.11 Enabling In-Depth Sentiment Analysis:

The thorough data preprocessing work made it possible to conduct a more thorough and insightful sentiment analysis of K-pop song lyrics. The sentiment analysis models were able to accurately interpret emotional nuances and pinpoint prevailing sentiments in the songs by creating a clear and organized dataset. In order to fully comprehend the emotional expressions conveyed through the alluring melodies of K-pop music, a thorough understanding of the data preprocessing stage's critical foundation was required.

The raw dataset was refined into a coherent and meaningful form through data preprocessing, which ultimately allowed for a thorough sentiment analysis of K-pop song lyrics. The dataset's dependability was ensured and made ready for precise sentiment analysis through the use of manual collection, web scraping, language translation, and data cleansing. These preprocessing efforts revealed insightful information about the emotional journey that K-pop songs take, illuminating the sentimental patterns that resound within this dynamic and well-liked musical genre.

Chapter 6

Data Exploration

Any data analysis or data science project must begin with data exploration. Investigating and comprehending the dataset is necessary to gain knowledge, spot patterns, and spot trends that can guide further analysis. The main goal of data exploration is to become familiar with the characteristics and relationships of the data so that one can make meaningful interpretations and defensible decisions.

6.1 Word Frequency Analysis

The exploratory data analysis of K-pop song lyrics uses word frequency analysis as a fundamental method. This method entails the extraction of frequently occurring words in order to spot recurring themes and gain understanding of the lyrical content. Analyzing the word frequencies allows for the identification of the words that recur most frequently in the songs as well as the interpretation of the meanings they express. The lyrics of K-pop songs were tokenized into separate words for this analysis, allowing us to carry out a thorough word frequency analysis. Common English stop words like "the," "and," and "is" were eliminated from the analysis to ensure its accuracy because they make little or no sense on their own.

After data processing, each word's occurrences were counted using the Python Counter class. The top ten words and their associated frequency are listed in the list below:

Like (2358): The word "like" appears to be the most frequently used word in the lyrics, indicating that similes and comparisons are frequently used in K-pop songs to convey emotions or experiences.

I'm (1816): "I'm" stands for both the first-person pronoun "I" and the verb "am." Given its high frequency, it suggests that K-pop performers frequently share their own emotions or ideas.

Love (1767): In K-pop lyrics, the word "love" is prominently used, emphasizing the theme of affection and emotional connection.

Know (1434): The word "know" suggests that the songs have a strong sense of self-awareness or a desire for comprehension.

Don't (1287): The contraction "don't" stand in for "do not" and frequently conveys a negation or resistance, giving the lyrical content more depth.

Go (1115): The word "go" may refer to motion, modification, or advancement in the songs' narratives.

Even (1066): The word "even" being present may denote contrast or surprise, hinting at intriguing lyrical turns.

Want (946): The word "want" suggests desires or aspirations that K-pop artists have expressed.

Heart (903): The word "heart" probably refers to the songs' passions, feelings, or deep sentiments.

One (856): The song titles all use the number "one" as a symbol for unity or singularity.

The word frequency analysis offers an overview of the key themes and feelings present in K-pop song lyrics. Words like "love," "know," and "heart" are frequently used, which suggests that emotional expression and self-awareness are common themes in K-pop music. This analysis lays the groundwork for more in-depth sentiment analysis and topic modeling in later stages of the study. It also helps us understand the lyrical content.

6.2 Sentiment Analysis

The key themes and emotions expressed in the lyrics of K-pop songs are outlined by the word frequency analysis. The frequent use of words like "love," "know," and "heart" suggests that emotional expression and self-awareness are prevalent themes in K-pop music. Later stages of the study will use this analysis to lay the groundwork for more thorough sentiment analysis and topic modeling. It also aids in our comprehension of the lyrics.

6.2.1 Sentiment Analysis Methodology

i. Lexicon-Based Sentiment Analysis with VADER: The sentiment analyzer had to be initialized after downloading the VADER lexicon. To ensure cleaner and more insightful results, the lyrics were tokenized and common stopwords, sexually explicit words, and special characters were eliminated.

ii. Calculating Compound Scores: The VADER analyzer created a compound score for each song that reflected the overall sentiment polarity. The scale for this score went from -1 (which represented the most extreme negative sentiment) to +1 (which represented the most extreme positive sentiment). Songs with compound scores above 0.5 were designated as "Positive," those with compound scores below -0.5 as "Negative," and those in the middle as "Neutral."

6.2.2 Sentiment Word Clouds

i.Positive Sentiment Word Cloud:

For each song, the VADER analyzer generated a compound score that represented the polarity of the overall sentiment. On a scale of -1 to +1, where -1 represented the most extreme negative sentiment and +1 the most extreme positive sentiment, this score was calculated. Songs with compound scores of 0.5 or more were labeled "Positive," those with compound scores of -0.5 or less were labeled "Negative," and those in the middle were labeled "Neutral."

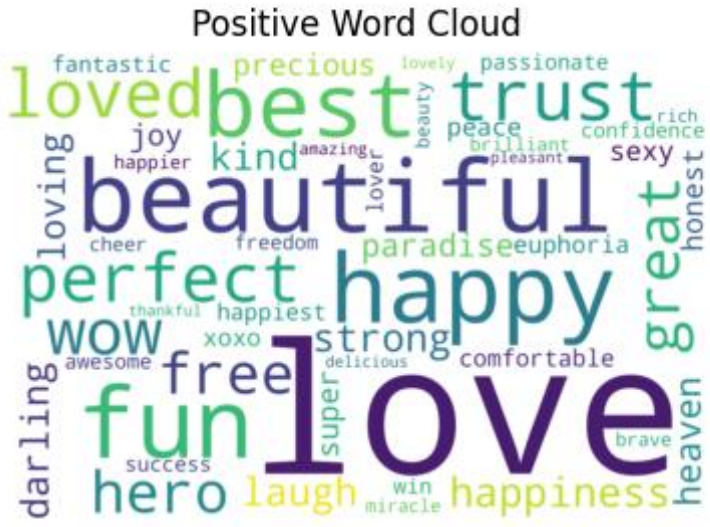

Figure 7. Positive Word Cloud

ii.Negative Sentiment Word Cloud: Words like "illegal," "bad," "pain," and "sick" were found in the word cloud for negative sentiment, providing insights into the infrequent consideration of gloomy themes in K-pop songs. Despite the predominance of upbeat and cheery melodies, K-pop also explores emotional and introspective themes, reflecting the complexity of human experiences. Words like "hurt," "hate," "rage," and "cruel" depict the darker emotions that are expressed in this genre. However, as evidenced by words like "die," which may denote resilience and rebirth, these songs also convey the strength to overcome obstacles.

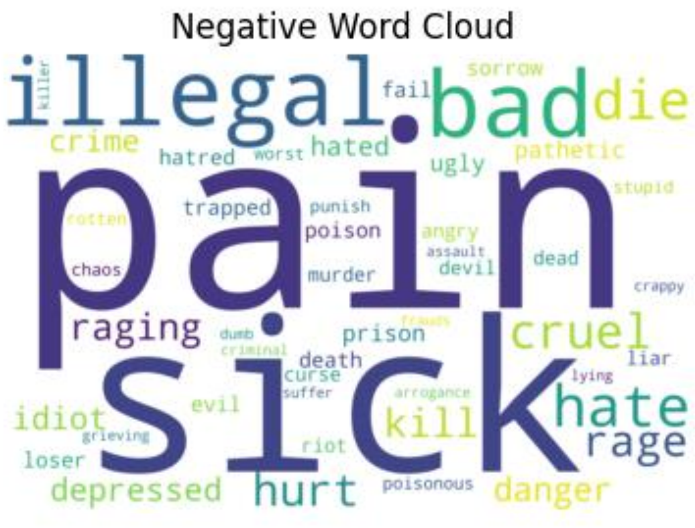
Negative Word Cloud

Figure 8. Negative Word Cloud

iii. Neutral Sentiment: Words like "heart," "game," "even," "scared," "butterfly," "come," "day," "time," "away," "look," and "see" were highlighted in the word cloud for neutral sentiment. Without explicit positive or negative connotations, these words imply moments of reflection, uncertainty, or observations. Neutrally sentimental K-pop songs offer a well-rounded viewpoint amidst the wide range of emotional themes that are common in the genre

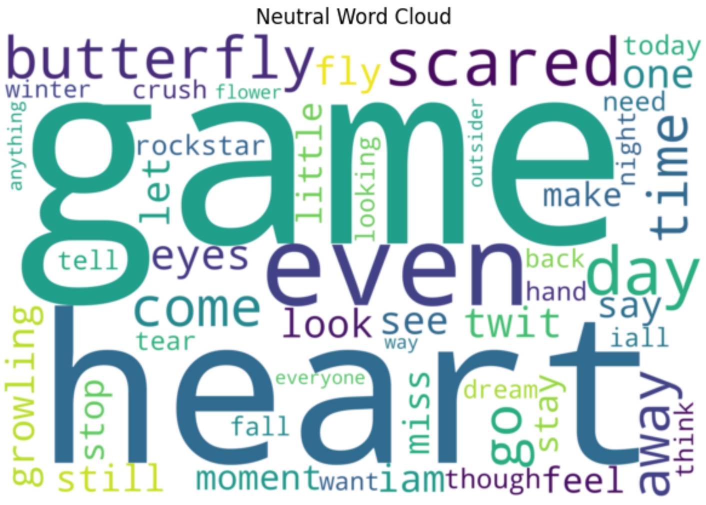
Neutral Word Cloud

Figure 9. Neutral Word Cloud

The exploration of the emotional landscape in K-pop songs through sentiment analysis was fascinating. Classifying songs according to their emotional content helps to better understand how a genre affects its audience. A complex picture of how K-pop artists create their work to connect with the wide range of emotions experienced by listeners around the world can be seen in the interaction of positive and negative emotions. Sentiment analysis and word clouds can be used to uncover the emotional web woven into K-pop music. K-pop transcends linguistic and cultural boundaries, eliciting a wide range of emotions in its listeners around the world, from joy and love to quiet moments of reflection. These results provide new directions for investigation and analysis, enhancing our understanding of the emotional depth that makes K-pop such a significant and cherished musical phenomenon.

6.3 Sentiment Distribution in K-pop Songs

The rich tapestry of emotions woven into K-pop songs is fascinatingly revealed by the sentiment distribution analysis. This analysis embarks on a profound exploration of the sentiments expressed within this genre, unravelling a spectrum of feelings that resonate with audiences all over the world using the labelled K-pop song dataset from "KpopMusic_Labelled.csv."The dataset is first imported into a Data Frame aptly named "data," which contains a crucial column titled "sentiment," before the analysis can begin. Each K-pop song is given a sentiment label in this column, indicating whether it is "Positive," "Negative," or "Neutral." This classification lays the groundwork for a thorough analysis of the emotional undercurrent that permeates the dataset.

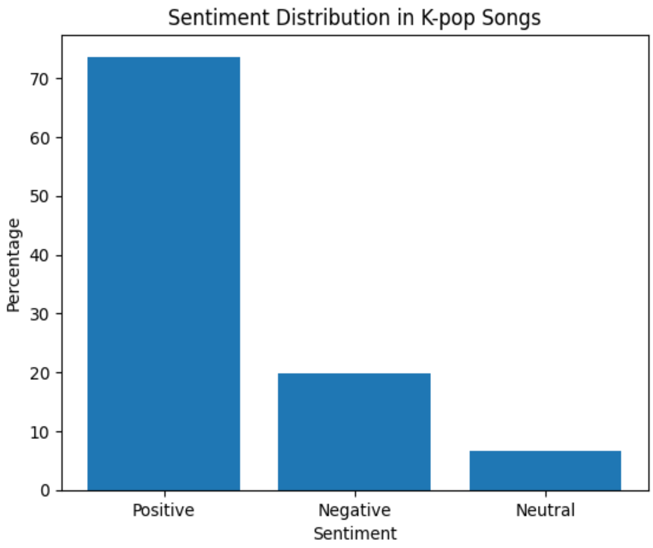

Figure 10. Sentiment Distribution in Kpop songs

After calculating the sentiment distribution, a visually striking bar plot of the emotional landscape of K-pop is produced. Three distinct bars, one for each of the percentages of "Positive," "Negative," and "Neutral" sentiments present in the dataset, are shown in the plot.

Positive Sentiment: "Positive" songs exude an overwhelming sense of optimism, love, and joy, standing tall as the prevailing sentiment. The prevalence of these inspiring themes demonstrates K-pop's capacity to inspire joy and inspiration in listeners around the world.

Negative Sentiment: The "Negative" sentiment category explores poignant and introspective themes, delving into feelings of heartbreak, pain, and struggle despite making up a smaller portion of the total. These songs demonstrate the versatility of the genre by exploring feelings that reflect the complexity of human experiences.

Neutral Sentiment: The analysis's notable finding is the substantial amount of "Neutral" sentiments. These songs travel a fascinating path between self-reflection, observation, and expressions that lack overtly positive or negative connotations. The presence of neutral emotions gives K-pop's emotional narrative more depth and allows for reflective moments amid the vibrant emotional tapestry.

The sentiment distribution analysis captures the essence of the feelings and experiences that K-pop songs are trying to express. This investigation broadens our understanding of the genre's capacity to arouse a range of emotions and engage a global audience. The bar plot's vivid depiction of the emotional diversity highlights it, reiterating K-pop's status as a significant and well-liked musical phenomenon.

6.4 Statistical Analysis

The essence of the emotions and experiences that K-pop songs are attempting to convey is captured by the sentiment distribution analysis. This study increases our knowledge of the genre's ability to elicit a variety of emotions and hold an audience on a global scale. K-pop's status as an important and well-liked musical phenomenon is reiterated by the bar plot's vivid depiction of the emotional diversity.

6.4.1 Converting Sentiment Scores to Numeric Values

The first step in the process is to convert the sentiment scores from their original string format of storage into a numeric format. The sentiment scores are translated to numerical values using Python's 'eval ()' function, laying the groundwork for further statistical calculations.

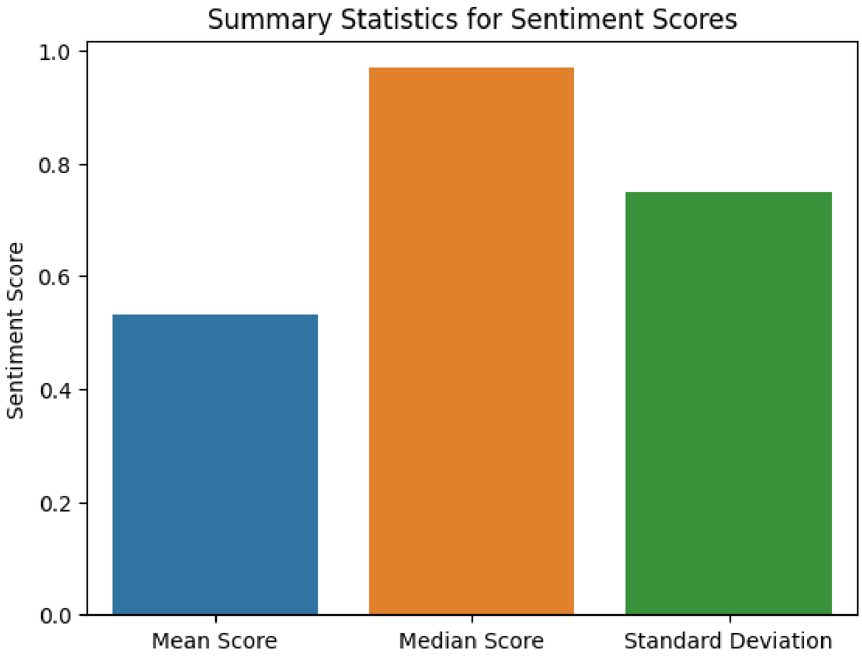

Figure 11. Summary Statistics for sentiment scores

The sentiment scores are now given as numerical values, and a thorough set of summary statistics is computed to give a clear picture of the emotional tendencies in the dataset.

Mean Score: The average sentiment conveyed by all K-pop songs is represented by the mean score, which was calculated as 0.526. This value, which is closer to the positive polarity end of the scale (which ranges from -1 to +1), indicates a primarily positive sentiment.

Median Score: The dataset is divided into two equal halves at the median score, which is calculated as 0.9648. This median score, which is close to 1, supports the generally upbeat sentiment present in K-pop songs.

Standard Deviation: The dispersion or variability of sentiment scores around the mean is represented by the standard deviation, which is calculated as 0.749. A higher standard deviation indicates that the dataset contains a wider range of expressed sentiments.

The high mean and median scores support the idea that K-pop songs primarily express positive emotions. The standard deviation of the dataset further denotes the coexistence of different emotional expressions, from highly positive to more neutral feelings. This variety highlights the genre's capacity to evoke a wide range of emotions in viewers and connect with them deeply. The statistical analysis reveals the emotional tapestry woven into K-pop music, reaffirming its ability to inspire listeners with love, joy, and positivity. Exploring these sentimental tendencies helps us understand the genre's profound influence on its worldwide fanbase.

6.5 Analyzing K-pop Song Topics with LDA (Latent Dirichlet Allocation)

K-pop songs frequently explore a wide range of emotions and themes thanks to their catchy melodies and captivating performances. Latent Dirichlet Allocation (LDA), a topic modeling technique, is used to determine the overarching themes in the lyrics of K-pop songs. Finding the main themes in the lyrics helps us understand the recurring themes and inventive expressions that characterize this dynamic musical genre [18].

6.5.1 Discovering Song Topics

The pre-processed lyrics were fed into the LDA model, which produced five different topics. A list of keywords that best capture the content of each topic is used to represent it.

i. Emotions and Personal Experiences: Words like "like," "know," "don't," "love," and "heart" indicate a focus on feelings and relationships. This topic is about emotions and personal experiences.

ii. Perseverance and Overcoming Challenges: The topic's keywords, such as "even," "run," "back," and "time," imply themes of tenacity and resolve, reflecting the K-pop aesthetic of overcoming challenges.

iii. Love and Desire: Words like "like," "love," "know," "get," and "want" indicate a romantic undertone and the desire for affection as this topic delves into themes of love and desire.

iv. Relationships and Experiences: Words like "love," "one," "night," and "day," which suggest topics about romantic encounters and the dynamics of love, are present in this topic, which is centred on relationships and experiences.

v. Personal Desires and Emotions: Words like "like," "don't," "want," "love," and "know" that refer to particular feelings and needs are used to highlight the topic's focus on individual needs and desires.

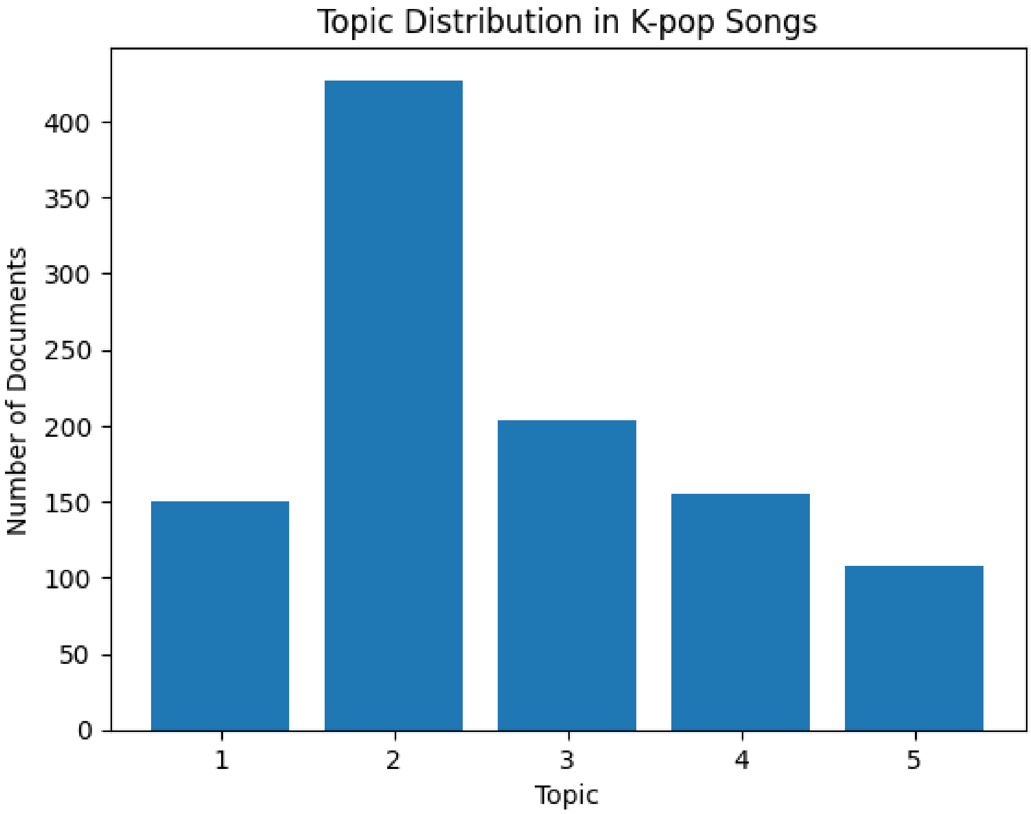

Figure 12. Visualization of LDA

The frequency of each identified topic across K-pop songs is shown in the topic distribution bar plot. Topic 2 comes out as the most common, indicating a significant amount of songs with themes of tenacity and overcoming obstacles. Topics 3, 1, and 4 are close behind, showing comparable frequencies. Last but not least, Topic 5 accounts for a relatively smaller percentage of songs in the dataset. The visualization provides a visual representation of the various subjects found in K-pop song lyrics, which helps us fully comprehend the thematic diversity and artistic expression of the genre.

The LDA-based topic modeling offers a thorough breakdown of the key themes found in K-pop songs. These issues strongly support the sentiment analysis's conclusions, reiterating the predominance of positive feelings and the expressiveness of K-pop music. Our understanding of the emotional and narrative elements communicated through K-pop song lyrics is deepened by the themes that have been identified, demonstrating the genre's artistic versatility and its capacity to connect with a wide audience. Overall, the topic modeling analysis deepens our understanding of the complex world of K-pop music by providing a window into the various feelings, narratives, and artistic expressions that support this dynamic musical phenomenon.

6.6 Comparison of K-pop and pop Music

Two incredibly popular and influential music genres that have won over music lovers all over the world are K-pop and pop music. These genres' comparisons are extremely valuable because it enables us to examine their shared influences, similar themes, and potential for cross-cultural expressions. Our appreciation of the similarities and differences between K-pop and Pop music enhances our understanding of their artistic development and influence on the global music scene.

K-pop incorporates elements from various cultures and musical genres, such as Pop and Hip-hop, to produce a distinctive and alluring sound. K-pop draws inspiration from a variety of musical genres, including Hip-hop and Pop. By contrasting K-pop and Pop music, the investigation enables the identification of the inspirations and influences that shape the distinctive characteristics of each genre. This makes it easier to comprehend the innovative ideas and tasteful fusion of musical genres that make K-pop such a vibrant and dynamic genre.

Even though K-pop and Pop music come from various cultural backgrounds, they frequently share emotional themes that connect with listeners all over the world. The lyrics' shared words and themes can be explored to learn more about how music can unite listeners from different linguistic and cultural backgrounds. Both K-pop and Pop songs recur with themes of love, friendship, dreams, and aspirations, demonstrating the profound power of music to evoke common human experiences.

In order to compare pop music and K-pop, a dataset of various pop music artists was downloaded from Kaggle and combined to form a new CSV file called "Pop_Songs.csv." Following that, this dataset was compared to the "KpopMusic_Labelled.csv" dataset of K-pop song lyrics. The goal of the analysis was to examine and comprehend the sentimental similarities and differences between these two genres, as well as their musical traits, in order to provide understanding of their emotional impact and overall significance [20].

6.6.1 Visualization of Commonly Used Words

The words most often used in K-pop and Pop song lyrics are shown in the bar graph below, along with words that are common to both genres. The goal of this visualization is to draw attention to the common word patterns and the ten words that appear most frequently in each genre.

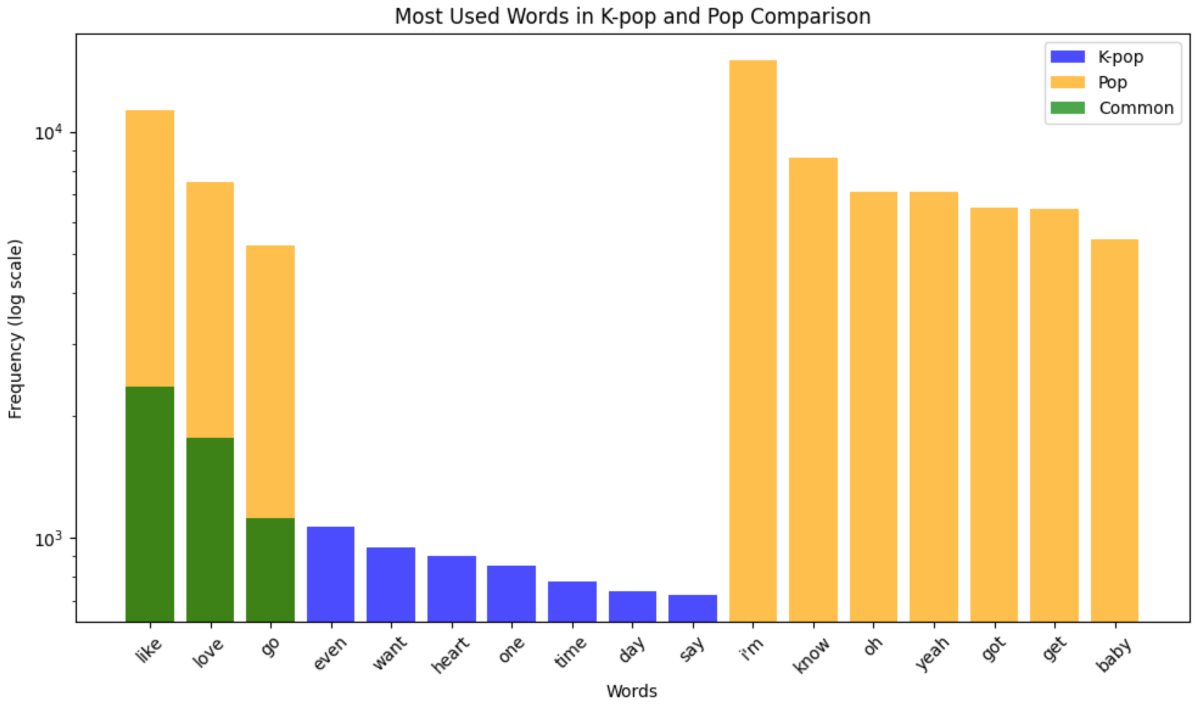

Figure 13. Most Used Words in K-pop and Pop Comparison

The most frequent words in K-pop music are represented by the blue bars in this visualization, whereas the most frequent words in pop music are represented by the orange bars. The words that are frequently used in both genres are highlighted by the green bars. The graph illustrates the frequency of words like "like," "love," and "go" in both K-pop and Pop music, highlighting the universal emotional themes that connect with listeners. Due to their emotional resonance and universal appeal, these words are frequently used. These basic human emotions and experiences are expressed in these straightforward yet potent words, which cut across linguistic and cultural barriers. Both genres hope to appeal to a wide international audience and evoke pleasant feelings by incorporating these well-known themes. This helps to promote a feeling of global unity and relatability among listeners. Additionally, the use of these words helps to overcome language barriers and contributes to the success of K-pop and Pop music on a global scale by making their music understandable and accessible to a wider range of fans.

While K-pop and Pop music may have some similar themes, they also celebrate their unique musical styles and individual identities. Through this comparison, Every genre of music is acknowledged for its unique contribution to the world of music, and the variety of music is respected. Exploring similarities and differences encourages fans to embrace the diversity of musical expressions and fosters a broad appreciation for musical creativity.

The comparison of K-pop and Pop music, in conclusion, highlights the harmonious fusion of artistic brilliance. Our knowledge of the larger musical world and its potential to unite people from various cultures and backgrounds is improved by exploring their shared influences, recurring themes, and universal appeal. K-pop and Pop music both possess the innate ability to move millions of people worldwide, despite having different histories and aesthetics. Not only are the lovely melodies and wide emotional range of these genres praised, but also the sense of global community that music fosters.

6.7 Title Sentiments in K-pop and Pop

Sentiment analysis can be used to investigate the emotional undertones in K-pop and Pop song titles. The sentiment of each title can be quantitatively assessed using the SentimentIntensityAnalyzer from NLTK, giving a better understanding of the general emotional orientation present in these genres.

6.7.1 Visualization of Average Title Sentiments

The average sentiment scores for K-pop and Pop music titles are displayed in the bar plot. Each score, which ranges from -1 (the most negative) to 1 (the most positive), represents the compound sentiment, which combines the polarity and intensity of emotions.

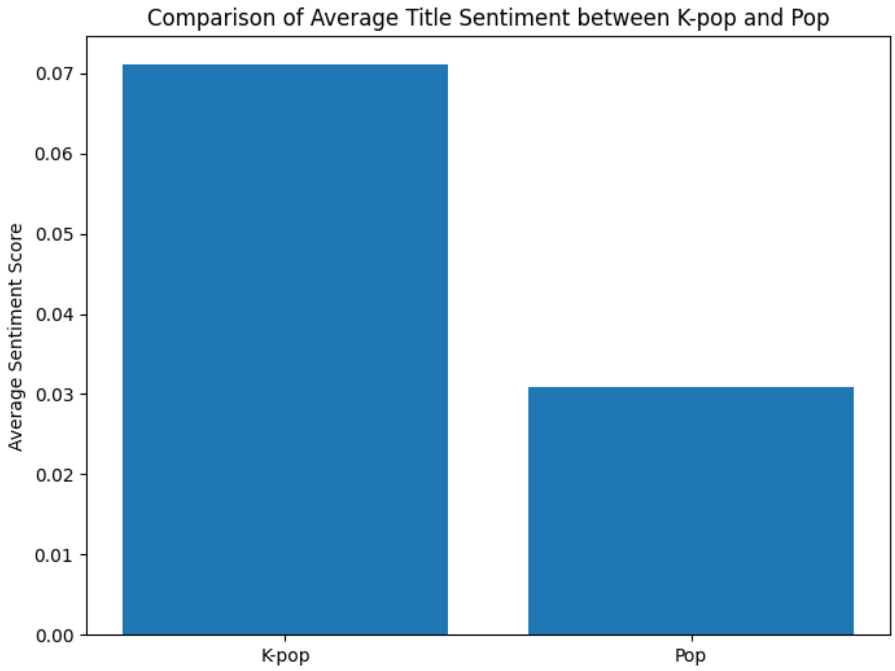

Figure 14. Comparison of Average Title Sentiment between K-pop and Pop

In the comparison, K-pop titles show a more upbeat sentiment than the typical title sentiment of Pop music. This finding suggests that K-pop titles frequently convey excitement, optimism, and positivity, in keeping with the exuberant and upbeat perception of the genre. The positive sentiment is also present in Pop music titles, despite having a slightly lower average, indicating that both genres contain themes and feelings that are upbeat.

6.8 Sentiment analysis vs. Danceability of K-pop Hits from 90's to 2020

An insightful analysis was carried out to explore the relationship between sentiment and danceability in K-pop hits spanning the years from the 1990s to 2020. For this study, two datasets were utilized: one featuring sentiment analysis labels for K-pop songs, and the other containing danceability scores for the same songs.

The sentiment analysis dataset was derived from the CSV file "KpopMusic_Labelled.csv," which encompassed song titles alongside their corresponding sentiment labels (Positive, Negative, or Neutral)

The song titles and associated danceability scores were contained in a CSV file called "merged_trend_file.csv," which served as the source for the danceability dataset used in this analysis. The data in this dataset illustrated the danceability traits of a wide range of K-pop songs from various decades, from the 1990s to the year 2020. This dataset was taken from a different Kaggle file that was specifically curated to include K-pop hits from various eras. Using data from this Kaggle dataset, the study examined danceability trends in K-pop music over time and discovered how danceability has evolved within the genre [19].

To facilitate the comparison, the datasets were merged based on the song titles using keyword search. Each song in the danceability dataset was matched with its corresponding title in the sentiment analysis dataset, allowing for the retrieval of sentiment labels. This process culminated in the creation of a consolidated dataset comprising song titles, danceability scores, and sentiment labels.

6.8.1 Visualization of Sentiment Vs Danceability

A scatter plot was created, with danceability on the x-axis and the sentiment on the y-axis. On the scatter plot, each data point represented a K-pop hit, and its colour denoted the sentiment label. Green represented songs with a positive message, red songs with a negative message, and blue songs with no message.

The scatter plot gave a visual representation of the correlation between sentiment and danceability in K-pop songs over time. Analyzing this plot can provide fascinating insights into whether there is a relationship between the emotional content of K-pop songs (as expressed by sentiment) and their danceability, a metric that determines a song's suitability for dancing.

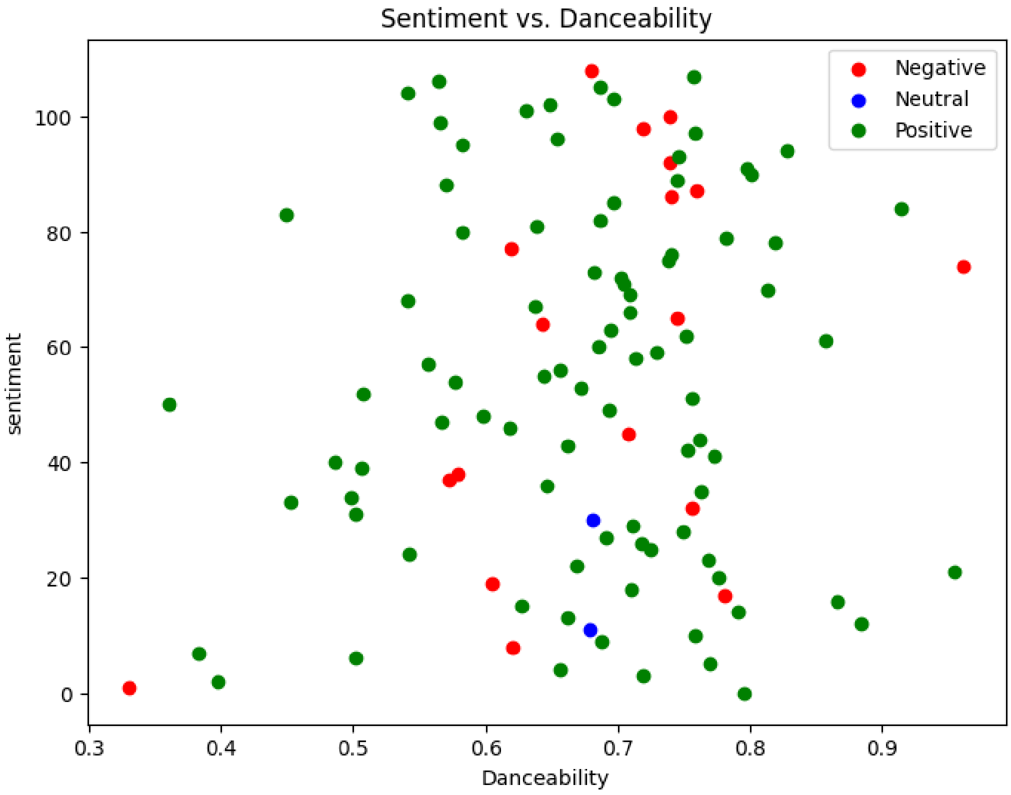

Figure 15. Sentiment vs Danceability

These results strongly support a relationship between sentiment and danceability in K-pop songs, indicating that a song's emotional content may affect how danceable it is. A song's danceability score may be higher if the lyrics express positive emotions; conversely, songs that express negative emotions may result in a composition that is less danceable. The existence of a relationship between sentiment and danceability may imply that emotional aspects of K-pop songs contribute to their danceability, potentially affecting how listeners relate to and react to the music on a deeper level.

6.8.2 Sentiment vs. Danceability: Summary Analysis

Summary analysis was conducted to explore the relationship between sentiment and danceability in the dataset of K-pop songs. The sentiment labels were categorized into three groups: Negative, Neutral, and Positive. Key statistical measures, including mean, median, standard deviation, and song count, were computed for each sentiment category to gain valuable insights into the danceability patterns associated with different emotional sentiments.

The analysis provides a glimpse into how danceability varies across songs with varying sentiment labels, shedding light on the potential correlation between emotional content and the danceability of K-pop songs.

Sentiment	Mean Danceability	Median Danceability	Standard Deviation	Song Count
Negative	0.683278	0.7135	0.127759	18
Neutral	0.680000	0.6800	0.001414	2
Positive	0.675045	0.6930	0.118026	89

Table 2. Sentiment vs. Danceability: Summary Analysis

The summary analysis reveals intriguing insights into the relationship between sentiment and danceability scores of K-pop songs. Notably, the mean danceability for songs with negative sentiment is marginally higher (0.683278) than that of songs with positive sentiment (0.675045). However, it is essential to consider the standard deviation, which indicates the dispersion of data points around the mean. The higher standard deviation for songs with negative sentiment (0.127759) compared to positive sentiment (0.118026) suggests that the danceability scores of songs with negative sentiment exhibit more variability.

Furthermore, the small sample size of songs with neutral sentiment (Song Count: 2) limits the generalization of observations for this sentiment category. Additional data points representing songs with neutral sentiment would be necessary for more robust conclusions.

Overall, this summary analysis provides valuable statistical insights, highlighting the subtle variations in danceability across different sentiment categories. However, further investigations using larger datasets would be instrumental in comprehensively understanding the intricate interplay between emotional content and danceability in the captivating world of K-pop songs.

Chapter 7

Implementation

This section discusses the application of supervised and unsupervised learning techniques for sentiment analysis on K-pop songs. The sentiments expressed in the lyrics are categorized using a variety of machine learning algorithms, including Logistic Regression, Naive Bayes, and Support Vector Machine (SVM), and their performance in sentiment prediction is compared. For thorough analysis, the lexicon-based VADER sentiment analyzer is also used. To find latent patterns and sentiments without labelled data, the potential of unsupervised learning using BERT (Bidirectional Encoder Representations from Transformers) is also investigated. This application is crucial to comprehending the emotional impact of K-pop songs on their global audience and provides insightful information about the wide range of emotional vistas depicted in this musical genre.

7.1 Supervised Learning Models

In this study, the sentiment analysis of K-pop songs involves the application of three supervised learning techniques: Logistic Regression, Naive Bayes, and Support Vector Machine (SVM). Each of these methods leverages labelled data to make predictions about the sentiments expressed in the lyrics. The goal is to classify the K-pop song lyrics into specific sentiment categories, including positive, negative, and neutral sentiments, by learning from the labelled data provided in the dataset. To label the data, the VADER sentiment analyzer is utilized, which assigns sentiment labels to the K-pop song lyrics. The VADER sentiment analyzer generates sentiment scores for each lyric in the dataset, determining whether the sentiment is positive, negative, or neutral. These sentiment labels are then used as the target labels for the supervised learning techniques employed in the sentiment analysis process. By harnessing the VADER sentiment analyzer for labelling the data, the models can effectively learn from the sentiment patterns and subsequently make accurate predictions on the sentiments expressed in the song lyrics.

The K-pop song lyrics dataset was pre-processed for the analysis by removing any inconsistencies and guaranteeing logical text representations. After pre-processing, the data was split into training and testing sets to make it easier to train and assess three classifiers: Logistic Regression, Naive Bayes, and Support Vector Machine (SVM).

The TF-IDF (Term Frequency-Inverse Document Frequency) technique was used to vectorize the text data in order to get the lyrics ready for model training. Through this procedure, each word in the lyrics was given a numerical representation that accounted for both its significance within the particular song and throughout the entire dataset.

By converting the textual data into numerical features, the classifiers will gain invaluable understanding and information that will enable them to make precise predictions about the emotions expressed in the song lyrics. The models will operate more quickly and efficiently during the sentiment analysis process as a result of this transformation, enabling them to effectively process numerical inputs. The precise sentiment analysis of K-pop song lyrics will eventually be made possible by pre-processing and TF-IDF vectorization, offering significant insights into the emotions conveyed through the music.

7.1.1 Logistic Regression:

As a statistical technique for binary outcomes, logistic regression is appropriate for sentiment analysis on K-pop song lyrics by predicting the sentiment expressed in the lyrics. The song lyrics and associated sentiment labels from the pre-processed data are fed into the Logistic Regression model. Before training, the lyrics go through TF-IDF vectorization, which turns them into numerical representations that take into account the significance of each word within the song and the entire dataset. The Logistic Regression model calculates the likelihood that a song's lyrics fall into the category of positive or negative sentiment during training by learning from the numerical features. The model is then used to predict the emotions expressed in each song's lyrics on the testing set. The effectiveness of the model is then assessed by comparing the predicted sentiments to the actual sentiment labels found in the testing set.

Numerous evaluation metrics, including accuracy, precision, recall, and F1-score, are computed to determine the effectiveness of the Logistic Regression model. Precision, recall, and F1-score give additional information about the model's capacity to distinguish between positive and negative emotions as well as the proportions of true positives, false positives, true negatives, and false negatives. Accuracy measures the percentage of correctly classified instances in the test set, while precision, recall, and F1-score give information about the capacity to distinguish between positive and negative emotions.

This sentiment analysis method seeks to offer insightful information into the emotions expressed in K-pop song lyrics, enabling a better understanding of the sentiment communicated through the music. It does this by using Logistic Regression as the classifier and feeding pre-processed data into the model.

7.1.2 Naive Bayes:

Due to its capacity to precisely predict the sentiment expressed in the text, the probabilistic classifier Naive Bayes is a highly effective technique. The Naive Bayes model is trained using the pre-processed song lyrics and their corresponding sentiment labels as input data. Naive Bayes learns the likelihood distribution of the lyrics' words and their correlation with favourable or unfavourable emotions during training. Naive Bayes determines the probability of a specific sentiment label for a song based on the occurrence of words in the lyrics by utilizing the conditional independence assumption. The sentiments for each song's lyrics were predicted using the model after training on the testing set. The actual sentiment labels found in the testing set were then contrasted with the predicted sentiments.

As with the evaluation of the Logistic Regression model, metrics like accuracy, precision, recall, and F1-score are used to assess the performance of the Naive Bayes classifier. The findings provided insight into the Naive Bayes classifier's overall efficacy in categorizing the sentiments expressed in K-pop song lyrics, as well as its capacity to distinguish between positive, negative and neutral sentiments.

7.1.3 Support Vector Machine (SVM):

For a variety of text analysis tasks, including sentiment analysis, Support Vector Machine (SVM) is a strong and adaptable classifier. It makes a great choice for predicting the sentiment expressed in K-pop song lyrics due to its capacity for handling high-dimensional data and finding the best hyperplanes for dividing various classes. The pre-processed song lyrics and their corresponding sentiment labels serve as input data for training the SVM model, much like Logistic Regression and Naive Bayes.

Next, the SVM model is trained using the transformed training data. During the training process, the model learns to establish a decision boundary, known as hyperplane, that effectively separates the sentiments. Following the training phase, the SVM model is applied to the testing set to predict the sentiments for each song's lyrics. The predicted sentiments are then compared against the actual sentiment labels present in the testing set.

Similar to the Naive Bayes classifier and logistic regression, the SVM classifier is evaluated by looking at important metrics like accuracy, precision, recall, and F1-score. A thorough understanding of the SVM classifier's performance in sentiment analysis tasks using K-pop songs can be obtained by analyzing these evaluation metrics. These metrics are crucial indicators of how well the classifier performs and whether it is appropriate for tasks involving sentiment analysis of K-pop songs.

7.2 Unsupervised Learning Models

In the context of sentiment analysis on K-pop song lyrics, unsupervised learning models play a crucial role in handling unlabelled data. These models do not rely on explicit target labels for training; instead, they deduce patterns and information from the data itself. In this study, two powerful unsupervised learning tools are utilized: Google BERT (Bidirectional Encoder Representations from Transformers) and VADER (Valence Aware Dictionary and sentiment Reasoner). Google BERT is a language representation model that encodes the context and semantics of words in a bidirectional manner, enabling it to capture intricate relationships and meaning within the text. On the other hand, VADER is a lexicon and rule-based sentiment analysis tool that derives sentiment scores based on the presence of sentiment-related words in the text. VADER's sentiment analysis is rule-based, relying on a pre-built lexicon that contains words and their associated sentiment scores. The sentiment labels for the unlabelled dataset can be approximated using VADER's rule-based method by utilizing the advantages of unsupervised models. These approximate sentiment labels, along with the fine-tuned BERT model, allow us to employ supervised learning techniques and achieve accurate sentiment analysis results on the K-pop song lyrics dataset.

7.2.1 BERT Model

This approach involves both supervised and unsupervised learning techniques. Initially, the VADER sentiment tool is utilized to generate sentiment scores for the training and validation sets. These sentiment scores play a crucial role in the subsequent fine-tuning process of the BERT model, a powerful language model.

During the unsupervised learning phase, the dataset's neutral sentiments are omitted from the training and validation sets. The decision to exclude neutral sentiments is primarily driven by the observation that there are relatively fewer songs with neutral sentiments compared to positive and negative sentiments. Consequently, including neutral sentiments in the unsupervised learning process might lead to an imbalanced dataset, potentially affecting the model's ability to generalize effectively to unseen data. By focusing solely on positive and negative sentiments during supervised learning, the model can better learn from the dominant sentiment patterns and improve its accuracy in predicting these two primary sentiment categories.

For the unsupervised learning part, the VADER algorithm is utilized to generate sentiment scores for the unlabelled dataset. However, due to the absence of explicit target labels, it is challenging to directly apply supervised learning techniques. Therefore, an approximation approach is employed, where sentiment labels are assigned based on the VADER-generated scores. Specifically, a sentiment label of 0 is assigned for negative sentiments if VADER's positive score is lower than the negative score, while a label of 1 is assigned for positive sentiments if VADER's positive score is higher than the negative score. Although this approach is not as precise as having manually labelled target sentiments, it enables sentiment approximation on the unlabelled dataset, making the unsupervised learning process feasible.

The BERT model is then fine-tuned using the training data, and the best-performing model is selected based on evaluation metrics. The model's performance is assessed on the test set by predicting sentiments for each lyric. By comparing the predicted sentiments with the manually obtained sentiment values in the test set, the overall accuracy, precision, recall, and F1 score of the BERT model in sentiment analysis can be evaluated.

By adopting this combined approach of supervised and unsupervised learning, the sentiment analysis of K-pop songs can be effectively conducted, providing valuable insights into the sentiments expressed in the lyrics. Successful sentiment analysis relies on the purposeful exclusion of neutral sentiments during supervised learning and the use of VADER-generated sentiment scores for unsupervised learning. Although the sentiment approximation used in unsupervised learning is less accurate than having labeled target sentiments, it still allows the model to make intelligent decisions based on the sentiment scores that are currently available.

Chapter 8

Evaluation and Results

8.1 Supervised methods

8.1.1 Logistic Regression

The evaluation metrics for the logistic regression model in sentiment analysis of K-pop songs are as follows:

Metric	**Value**
Accuracy	0.772
Precision	0.791
Recall	0.772
F1Score	0.702

Table 3. Evaluation metrics for the logistic regression model

The logistic regression model achieved favourable discriminative performance for classifying sentiment in a multi-class classification task. The accuracy of 0.772 indicates that approximately 77.2% of the instances were correctly classified, demonstrating the model's effectiveness in distinguishing between positive, negative, and neutral sentiments.

The precision of 0.791 implies that among the instances predicted as positive sentiment, around 79.1% were truly positive. The recall of 0.772 indicates that the model correctly identified approximately 77.2% of the actual positive sentiment instances. The F1 score of 0.702 represents the harmonic mean of precision and recall, providing an overall measure of the model's ability to correctly identify positive sentiment instances while minimizing false positives.

Confusion Matrix:

The confusion matrix for the logistic regression model is as follows:

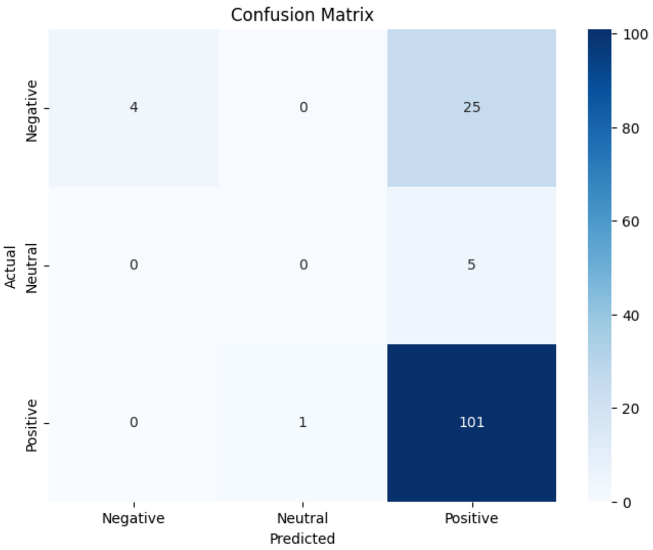

Figure 16. Confusion matrix of Logistic Regression Model

The results of the confusion matrix for the Logistic Regression model are as follows:

True Negatives (TN): Four instances of negative sentiment were correctly classified as such by the model.

101 instances of positive sentiments were correctly predicted as positive by the model, making them true positives (TP).

False Negatives (FN): In one instance, the model incorrectly categorized a positive sentiment as a negative one.

Five instances of neutral sentiment were mistakenly categorized as positive by the model, leading to false positives (FP).

As evidenced by the high counts of TN and TP, the Logistic Regression model successfully identified both negative and positive sentiments. FP misclassifications occurred as a result of difficulties it ran into when dealing with neutral sentiments.

ROC Curve Results:

The ROC curve for Logistic Regression model is as follows:

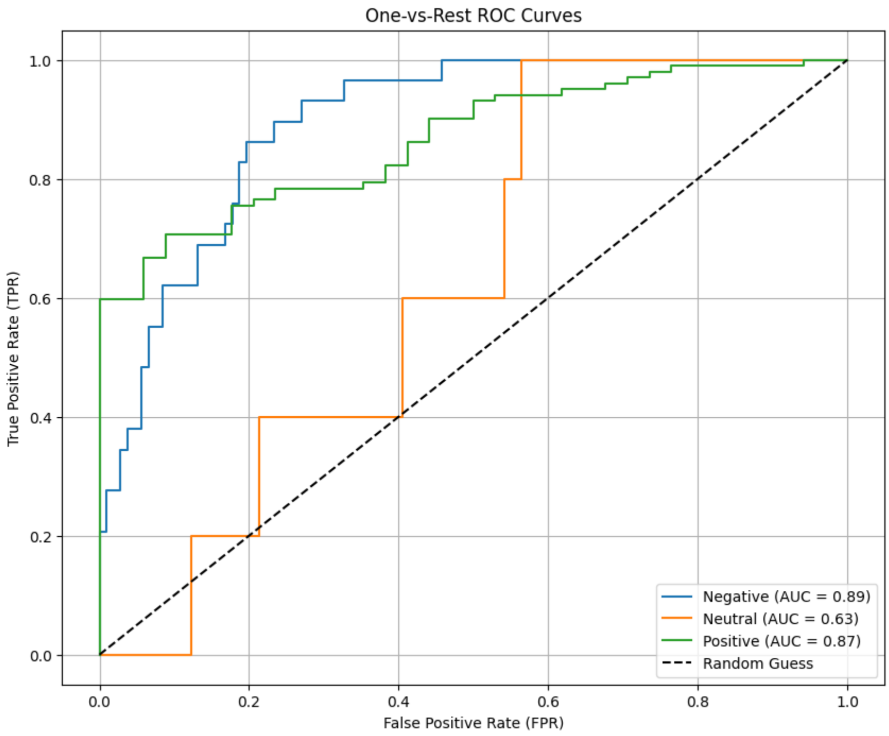

Figure 17. ROC curve of Logistic Regression Model

Negative Class (AUC = 0.89): The AUC score of 0.89 indicates that the model performs very well in distinguishing negative sentiment instances from other sentiments. The ROC curve for the negative class is situated closer to the top-left corner, indicating high true positive rates and low false positive rates for the negative sentiment class.

Neutral Class (AUC = 0.63): The AUC score of 0.63 suggests that the model's performance for classifying neutral sentiment instances is relatively moderate. The ROC curve for the neutral class might not be as close to the top-left corner as desired, indicating that the model's ability to differentiate neutral sentiment from other sentiments is not as strong as for the negative and positive classes.

Positive Class (AUC = 0.87): The AUC score of 0.87 indicates that the model performs well in distinguishing positive sentiment instances from other sentiments. The ROC curve for the positive class is closer to the top-left corner, indicating high true positive rates and low false positive rates for the positive sentiment class.

The logistic regression model demonstrated strong abilities in distinguishing negative and positive sentiments, with AUC scores of 0.89 and 0.87, respectively, as indicated by the ROC curves situated closer to the top-left corner. However, its performance in identifying neutral sentiment was comparatively moderate (AUC = 0.63), with the ROC curve not as close to the ideal corner. Despite this, the model's overall performance surpassed random guessing, evident by the ROC curves lying above the random guess line. The evaluation metrics and analysis provide insights into the model's effectiveness in sentiment classification for K-pop song lyrics.

8.1.2 Naive Bayes

The evaluation metrics for the Naive Bayes model in sentiment analysis of K-pop songs are as follows:

Metric	**Value**
Accuracy	0.772
Precision	0.791
Recall	0.772
F1Score	0.702

Table 4. Evaluation metrics for the Naive Bayes model

The Naive Bayes model achieved moderate performance in classifying sentiment in a multi-class classification task. The accuracy of 0.75 indicates that approximately 75% of the instances were correctly classified, demonstrating the model's ability to distinguish between positive, negative, and neutral sentiments.

The precision of 0.5625 implies that among the instances predicted as positive sentiment, around 56.25% were truly positive. The recall of 0.75 indicates that the model correctly identified approximately 75% of the actual positive sentiment instances. The F1 score of 0.6429 represents the harmonic mean of precision and recall, providing an overall measure of the model's ability to correctly identify positive sentiment instances while minimizing false positives.

Confusion Matrix:

The confusion matrix for the Naive Bayes model is as follows:

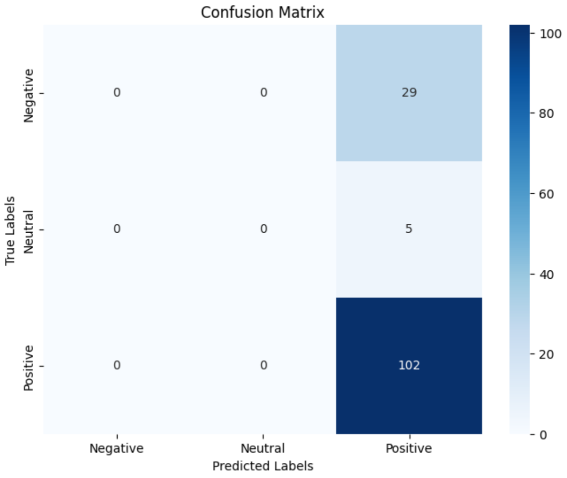

Figure 18 Confusion matrix of Naive Bayes model

These findings are drawn from the confusion matrix:

True Negatives (TN): The Naive Bayes model correctly identified 29 instances of negative sentiment as negative.

True Positives (TP): 102 instances of Positive sentiments were correctly predicted by the model to be Positive.

False Negatives (FN): The model never mistakenly classified negative sentiments as positive or neutral, showing that this error never occurred.

False Positives (FP): The model misclassified five instances of neutral sentiment as positive.

However, as evidenced by the existence of False Positives (FP), the model encountered difficulties when addressing Neutral sentiments. This indicates a potential area for model improvement to increase the model's accuracy in predicting Neutral sentiments. Some Neutral instances were mistakenly classified as Positive in this situation.

ROC Curve Results:

The ROC curve for Naive Bayes model is as follows:

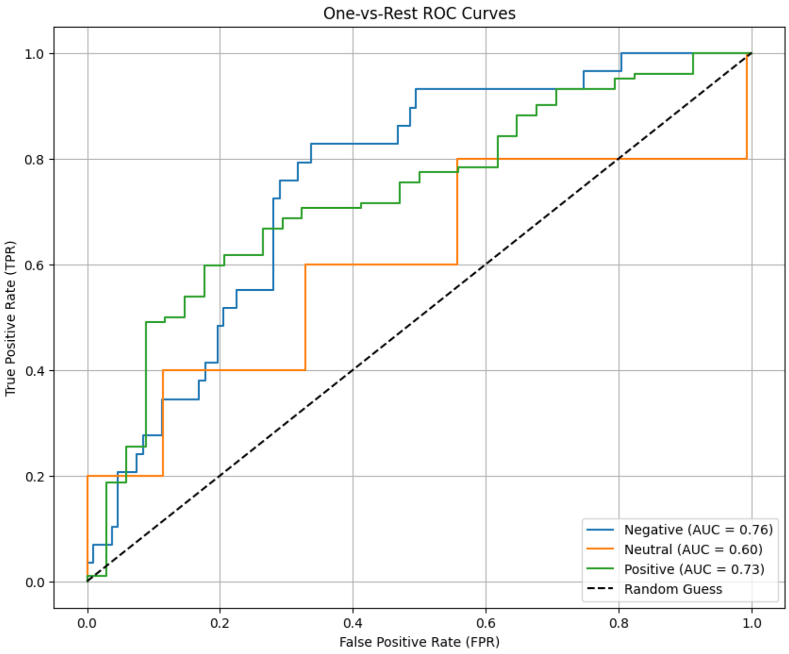

Figure 19. ROC Curve of Naive Bayes model

Negative Class (AUC = 0.76): The AUC score of 0.76 indicates that the model performs well in distinguishing negative sentiment instances from other sentiments. The ROC curve for the negative class is situated closer to the top-left corner, indicating relatively high true positive rates and low false positive rates for the negative sentiment class.

Neutral Class (AUC = 0.60): The AUC score of 0.60 suggests that the model's performance for classifying neutral sentiment instances is relatively moderate. The ROC curve for the neutral class might not be as close to the top-left corner as desired, indicating that the model's ability to differentiate neutral sentiment from other sentiments is not as strong as for the negative and positive classes.

Positive Class (AUC = 0.73): The AUC score of 0.73 indicates that the model performs reasonably well in distinguishing positive sentiment instances from other sentiments. The ROC curve for the positive class is closer to the top-left corner, indicating moderate true positive rates and low false positive rates for the positive sentiment class.

The Naive Bayes model demonstrated moderate performance in distinguishing positive and negative sentiments, with AUC scores of 0.73 and 0.76, respectively, as indicated by the ROC curves situated closer to the top-left corner. However, its performance in identifying neutral sentiment was relatively weaker (AUC = 0.60), with the ROC curve not as close to the ideal corner. Despite this, the model's overall performance surpassed random guessing, evident by the ROC curves lying above the random guess line.

8.1.3 Support Vector Machines

The evaluation metrics for the Support Vector Machines in sentiment analysis of K-pop songs are as follows:

Metric	**Value**
Accuracy	0.779
Precision	0.796
Recall	0.779
F1Score	0.716

Table 5. Evaluation metrics for the Support Vector Machines

The Support Vector Machines model demonstrated favourable performance in sentiment classification for K-pop song lyrics. With an accuracy of 0.779, the model correctly classified approximately 77.9% of the instances, showcasing its effectiveness in distinguishing between positive, negative, and neutral sentiments. The precision of 0.796 implies that among the instances predicted as positive sentiment, around 79.6% were truly positive. The recall of 0.779 indicates that the model correctly identified approximately 77.9% of the actual positive sentiment instances. The F1 score of 0.716 represents the harmonic mean of precision and recall, providing an overall measure of the model's ability to correctly identify positive sentiment instances while minimizing false positives.

Confusion Matrix:

The confusion matrix for the Support Vector Machines model is as follows:

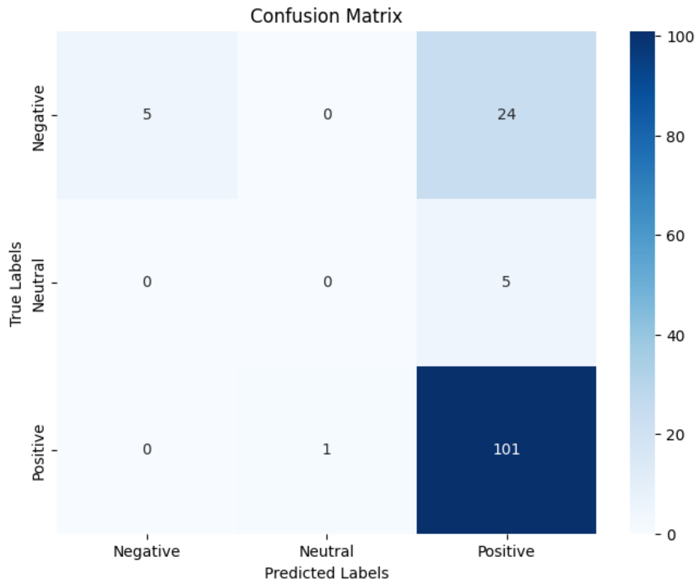

Figure 20. Confusion matrix of SVM Model

The confusion matrix for the SVM model yielded the following conclusions:

- True Negatives (TN): The SVM model correctly classified 24 instances of negative sentiment as negative.

- True Positives (TP): The model correctly identified 101 instances of positive sentiment as positive.

- False Negatives (FN): The model incorrectly categorized one instance of positive sentiment as negative.

- False Positives (FP): The model incorrectly categorized five instances of neutral sentiment as positive.

The SVM model correctly identified both negative and positive sentiments with a satisfactory level of accuracy, as shown by the high counts of TN and TP. The model, though, had trouble telling apart neutral emotions, which led to FP misclassifications. This points to a potential area for model enhancement to boost the model's capability of foretelling neutral sentiments. The effectiveness of the model as a whole can be better understood through additional research and consideration of various evaluation metrics, which can also suggest improvements for tasks involving sentiment analysis.

ROC Curve Results:

The ROC curve for SVM model is as follows:

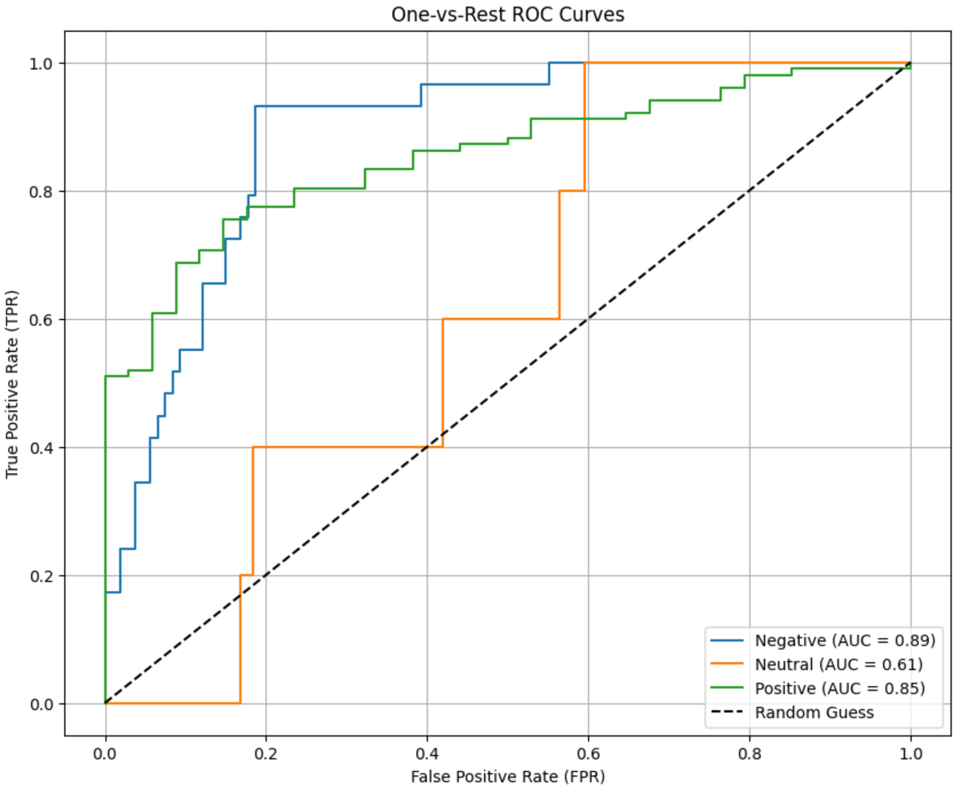

Figure 21. ROC curve of SVM Model

Negative Class (AUC = 0.89): The AUC score of 0.89 indicates that the model performs very well in distinguishing negative sentiment instances from other sentiments.

Neutral Class (AUC = 0.60): The AUC score of 0.60 suggests that the model's performance for classifying neutral sentiment instances is relatively moderate.

Positive Class (AUC = 0.85): The AUC score of 0.85 indicates that the model performs well in distinguishing positive sentiment instances from other sentiments.

The Naive Bayes model demonstrated strong abilities in distinguishing negative and positive sentiments, with AUC scores of 0.89 and 0.85, respectively, as indicated by the ROC curves. However, its performance in identifying neutral sentiment was comparatively moderate (AUC = 0.60). The evaluation metrics and analysis provide insights into the model's effectiveness in sentiment classification for K-pop song lyrics.

8.1.4 Comparing Sentiment Analysis Models

This section compares the effectiveness of Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) sentiment analysis models on K-pop song lyrics. A number of metrics, such as accuracy, precision, recall, and F1 score, are used to evaluate each model.

Model	**Accuracy**	**Precision**	**Recall**	**F1Score**
Logistic Regression	0.772	0.791	0.772	0.702
Naïve Bayes	0.750	0.796	0.750	0.716
SVM	0.779	0.796	0.779	0.716

Table 6. Comparison of Sentiment Analysis Models

From the table above, the SVM model outperformed the Logistic Regression model (0.772) and the Naive Bayes model (0.750), achieving the highest accuracy (0.779). But it's also crucial to take other metrics into account.

When it comes to precision, which indicates the proportion of true positive instances among all predicted positive instances, the Naive Bayes model achieved the highest precision (0.796), followed closely by both the Logistic Regression (0.791) and SVM (0.796) models.

In terms of recall, which represents the proportion of true positive instances among all actual positive instances, the Logistic Regression model achieved the highest recall (0.772), followed by the SVM model (0.779), and the Naive Bayes model (0.750).

The F1 score, which is the harmonic mean of precision and recall, provides a balanced measure of the model's overall performance. The SVM model achieved the highest F1 score (0.716), followed by the Naive Bayes model (0.716), and the Logistic Regression model (0.702).

It's important to note that there was a problem with the distribution of neutral sentiment instances when analyzing the sentiment of K-pop song lyrics. The dataset contained a relatively small number of songs with neutral sentiments compared to positive and negative sentiments. As a result, the models had fewer neutral instances to learn from during training, which might have affected their performance in classifying neutral sentiments accurately.

Furthermore, it is important to consider the labelling method used for the sentiment analysis. In this study, the lyrics were labelled using the VADER sentiment analyzer, which assigns sentiment scores to text based on the presence of sentiment-related words. As a rule-based sentiment analysis tool, VADER provides a continuous sentiment score rather than discrete labels. This approach offers advantages in capturing fine-grained sentiment information. Consequently, the probability of achieving high accuracy in this case is increased as it considers a wide range of sentiment expressions in the lyrics.

In conclusion, the comparison of the sentiment analysis models, along with the consideration of the distribution of neutral sentiments and the use of VADER for labelling, provides valuable insights into their respective strengths and weaknesses. These findings can aid in selecting the most appropriate model for sentiment analysis tasks in the context of K-pop songs.

8.2 Unsupervised methods

8.2.1 VADER – BERT Base approach

The "VADER – BERT Base Approach - Epoch Results" provides a comprehensive analysis of the model's training performance over multiple epochs. The table below displays presents the performance comparison of the VADER – BERT Base unsupervised approach for sentiment analysis of K-pop song lyrics across different epochs. The table above showcases the accuracy, precision, recall, and F1 score for each epoch.

Epoch	Accuracy	Precision	Recall	F1Score
5	0.7743	0.6674	0.7743	0.6961
10	0.7874	0.6200	0.7874	0.6937
15	0.7874	0.6200	0.7874	0.6937
20	0.8189	0.8289	0.8189	0.7671

Table 7. Comparing VADER and BERT Base Approaches - Epochs

1. Epoch 5:

At epoch 5, the model exhibits an accuracy of 77.43% and achieves a balanced F1 Score of 69.61%. The precision and recall are moderately balanced, suggesting that the model is making accurate positive and negative predictions, though some improvements are possible.

2. Epoch 10:

At epoch 10, the model shows an accuracy of 78.74%. However, the precision drops to 62.00%, indicating that the positive predictions are not as accurate as in the previous epoch. The F1 Score remains similar to epoch 5, suggesting that the model's overall performance has not improved significantly.

3. Epoch 15:

At epoch 15, the model's performance remains consistent with epoch 10, with an accuracy of 78.74% and a precision of 62.00%. The recall and F1 Score are also the same, indicating that the model's ability to identify positive instances and its overall performance have not changed.

4. Epoch 20 (Best Checkpoint):

At epoch 20, the model shows notable improvements in all metrics. It achieves the highest accuracy of 81.89% among all epochs and demonstrates a precision of 82.89%, indicating a significant boost in correctly identifying positive instances. The F1 Score also increases, reaching 76.71%, suggesting a more balanced trade-off between precision and recall.

Best Performing Epoch:

Epoch 20 stands out as the best-performing checkpoint for sentiment analysis. It exhibits the highest accuracy, precision, recall, and F1 score, showcasing its capability to generalize well to unseen data and provide accurate sentiment predictions.

Best Checkpoint Prediction Analysis:

Using the model checkpoint from **Epoch 20**, the sentiment analysis was performed on a new set of K-pop song lyrics. The predictions yielded the following results:

Metric	**Value**
Accuracy	0.78125
Precision	0.775
Recall	0.9894
F1Score	0.8692

Table 8. Results of Model checkpoint from Epoch 20

The model demonstrated high precision (0.775), indicating its ability to accurately identify positive instances. It also achieved an impressively high recall (0.9894), suggesting that it can effectively identify the majority of positive cases. The F1 Score (0.8692) indicates a balanced trade-off between precision and recall.

The analysis of the VADER – BERT Base unsupervised approach for sentiment analysis at different epochs shows improvements in performance metrics as training progresses. Among the epochs, Epoch 20 proves to be the best checkpoint for sentiment analysis, with the highest accuracy, precision, recall, and F1 score. Utilizing Epoch 20 as the optimal checkpoint for prediction, the model exhibited impressive precision and recall on new data, enabling accurate sentiment analysis of K-pop song lyrics. With this method, it is possible to comprehend and appreciate the emotional expressions that are conveyed in this popular musical genre on a deeper level. Epoch 20 can be used with confidence to predict sentiments in new and varied K-pop songs, giving important insights into the sentiments and feelings expressed in the lyrics.

Chapter 9

Conclusion

Sentiment analysis for K-pop songs was conducted using a dataset of 680 songs, comprising song titles and lyrics. The journey encompassed various stages, starting with meticulous data preprocessing, which involved the removal of stop words, explicit words, and cleaning of the lyrics text to establish a robust foundation for analysis.

During the exploratory data analysis (EDA) phase, valuable insights were gained into word frequencies and sentiment distributions within K-pop songs. Word cloud visualizations vividly portrayed the usage of positive and negative words, providing a glimpse into the emotional content conveyed through the lyrics. Additionally, a comparison was conducted between sentiment analysis and danceability scores of songs. The scatter plot visualization revealed that songs with positive sentiment tend to have higher danceability scores, forming a distinct cluster in the upper range. On the other hand, songs with negative sentiment are associated with lower danceability scores, forming a separate cluster in the lower range. Neutral sentiment songs were few and showed no clear pattern in danceability. These findings suggest a correlation between sentiment and danceability, implying that emotional content may influence the danceability of songs.

The application of VADER sentiment analysis successfully labelled the song data with sentiment labels, enabling a deeper analysis of the emotional sentiments present in K-pop songs. For predictive modeling, three supervised approaches were explored: logistic regression, SVM, and Naive Bayes. These models demonstrated their capabilities in predicting sentiments for K-pop songs, achieving accuracy scores of 77.2%, 77.9%, and 75.0%, respectively. A comprehensive evaluation of various sentiment analysis models, including VADER Sentiment Analyzer, Logistic Regression, Naive Bayes, and SVM, was performed, providing a thorough understanding of their strengths and limitations.

One of the project's highlights was the exploration of unsupervised training using BERT, a powerful transformer-based approach. By fine-tuning BERT on labelled data, an impressive accuracy of 78.125% was achieved, outperforming other supervised approaches.

In conclusion, the findings from this sentiment analysis of K-pop songs provide valuable insights into the emotional expressions embedded in the song lyrics. The powerful transformer-based approach using BERT showcased a remarkable accuracy, highlighting its effectiveness in sentiment analysis for cases with limited labelled data. The comparison with danceability scores further deepens our understanding of the relationship between sentiment and the musical attributes of songs. These results contribute significantly to the understanding of sentiment analysis in the context of music, particularly K-pop songs, shedding light on the sentiments conveyed through this popular genre. Researchers, music enthusiasts, and industry professionals can leverage this knowledge to better comprehend the emotional impact of K-pop songs and their profound connection with the audience.

9.1 FUTURE WORKS

This project's conclusion reveals several vital directions for future work on sentiment analysis for K-pop songs, allowing for a more thorough examination of the emotions expressed in musical compositions. The limited supply of labelled data needed to train sentiment analysis models was a significant obstacle in this project. Future research should concentrate on creating more comprehensive and varied labelled datasets in order to improve the accuracy and generalization of sentiment analysis models. To enable more robust model training, this may entail making an effort to annotate larger sets of K-pop songs with sentiment labels.

Additional metadata, such as the song's release year, the artist's biography, and the type of music, could add significant context to sentiment analysis. By utilizing such metadata, it is possible to analyze sentiment trends over time, compare feelings between artists or genres, and comprehend the influence of different factors on song sentiment.

Sentiment analysis of K-pop songs can be significantly impacted by taking into account the context of their creation and release. Understanding the emotions expressed in the songs more fully can be achieved by taking into account elements like the artist's personal experiences, historical events, or cultural context.

A user-centric perspective on the emotional reception of K-pop songs may be obtained by combining sentiment analysis with user feedback and engagement data. Music producers and marketers could benefit from understanding how listeners interpret and react emotionally to particular songs.

Given BERT's performance in this project, it may be possible to use more extensive pre-trained language models in the future, such as bigger versions of BERT or other cutting-edge transformer-based models. Accurate sentiment analysis may be further improved by fine-tuning these models on larger datasets.

The conclusion of this project demonstrates exciting possibilities for future study in sentiment analysis for K-pop music. Sentiment analysis methods can be improved in order to increase our comprehension of the emotional expressions expressed through this popular music genre of K-pop by addressing the problems of data scarcity, incorporating additional song information, taking context into account, and investigating various types of K-pop music. As sentiment analysis develops, it will undoubtedly help us understand music's ability to evoke strong emotions in listeners all over the world.

References:

[1] Hayeon Jang, Munhyong Kim, and Hyopil Shin, "KOSAC: A Full-fledged Korean Sentiment Analysis Corpus," Pacific Asia Conference on Language, Information and Computation, 2013.

[2] Humberto Corona and Michael P. O'Mahony, "An Exploration of Mood Classification in the Million Songs Dataset," 12th Sound and Music Computing Conference at Maynooth, Ireland, 2015.

[3] Yeawon Yoo and Yonghan Ju, "Quantitative analysis of a half-century of K-Pop songs: Association rule analysis of lyrics and social network analysis of singers and composers," Journal of Popular Music Studies 29, 2017.

[4] Shivaji Alaparthi and Manit Mishra, "Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey", 2020.

[5] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, Foundations and Trends® in Information Retrieval 2(1–2):1-135, 2008.

[6] Ronen Feldman, "Techniques and applications for sentiment analysis," Communications of the ACM 56(4): 82–89, 2013.

[7] Karthick Prasad Gunasekaran, "Exploring Sentiment Analysis Techniques in Natural Language Processing: A Comprehensive Review", 2023.

[8] C.J. Hutto and Eric Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," Conference: Proceedings of the Eighth International AAAI Conference on Weblogs and social media at Ann Arbor, MI, 2015.

[9] Nonita Sharma, Monika Mangla and Sachi Nandan Mohanty, "Supervised Learning Techniques for Sentiment Analysis," Emerging Technologies in Data Mining and Information Security (pp.423-435), 2022

[10] Kahyun Choi, Jin Ha Lee and Craig Willis J. Stephen Downie, "Topic Modeling Users Interpretations of Songs to Inform Subject Access in Music Digital Libraries," the 15th ACM/IEEE-CE, 2015.

[11] Katleen Napier, Lior Shamir, "Quantitative Sentiment Analysis of Lyrics in Popular Music," Journal of Popular Music Studies 30(4):161-176, 2018

[12] Joon-ho Kim,Seung-hye Jung,Jung-sik Roh and Hyun-ju Choi 4, "Success Factors and Sustainability of the K-Pop Industry: A Structural Equation Model and Fuzzy Set Analysis," Sustainability 2021, 13(11), 5927, 2021.

[13] Kaggle, "BTS Lyrics Dataset," [Online]. Available: https://www.kaggle.com/dataset/bts-lyrics. [Accessed: April 4, 2023].

[14] KPop Lyrics, " KPop Lyrics - Translations & Romanizations," [Online]. Available: https://www.kpoplyrics.net/. [Accessed: April 15, 2023].

[15] Towards Data Science, " A Short Introduction to VADER," [Online]. Available https://towardsdatascience.com/an-short-introduction-to-vader-3f3860208d53. [Accessed: April 20, 2023].

[16] observablehq, "Analyzing popular K-pop song lyrics," [Online]. Available: https://observablehq.com/@71/analyzing-popular-k-pop-song-lyrics. [Accessed: May 2, 2023].

[17] Analytics Vidhya, " 12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2023)," [Online]. Available: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics. [Accessed: June 10, 2023].

[18] Medium, " What’s in a Song? LDA Topic Modeling of over 120,000 Lyrics," [Online]. Available https://tim-denzler.medium.com/whats-in-a-song-using-lda-to-find-topics-in-over-120-000-songs-53785767b692. [Accessed: June 20, 2023].

[19] Kaggle, " General_Trend_In_KPop," [Online]. Available: https://www.kaggle.com dataset/bts-lyrics. [Accessed: June 15, 2023].

[20] Kaggle, " Song Lyrics Dataset - A dataset containing song lyrics of various artists” [Online]. Available: https://www.kaggle.com/datasets/deepshah16/song-lyrics-dataset. [Accessed: June 25, 2023].

[21] Hugging Face, "Bert-base-multilingual-uncased," [Online]. Available https://huggingface.co/bert-base-multilingual-uncased. [Accessed: June 28, 2023].