ISSN 2085-4552

Sentiment Analysis on Song Lyrics for Song Popularity Prediction Using BERT Algorithm

Hana Agatha[1], Farica Perdana Putri[2], Alethea Suryadibrata[3]
Department of Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia
[1]hana.agatha@student.umn.ac.id, [2]farica@umn.ac.id, [3]alethea.suryadibrata@lecturer.umn.ac.id

Accepted 29 November 2023
Approved 20 December 2023

***Abstract*—The increasing competitiveness in the music industry is giving some musicians disadvantages. Musicians need to pay more attention to the factors that influence the popularity of a song, so their song can be popular and they can gain a lot of profit. One of the various factors that can affect a song's popularity is the lyrics. The influence of the lyrics can be explored through sentiment analysis. Sentiment analysis is a computing study that identifies sentiments or emotions in a text. By conducting sentiment analysis on the lyrics, song popularity can be predicted. Based on the prediction result, songwriters can evaluate their lyrics, so their songs can be popular. Bidirectional Encoder Representations from Transformers (BERT) is an excellent algorithm in terms of sentiment analysis. In this study, a BERT model was developed to predict the song's popularity, based on the sentiment analysis of the song lyrics. The popularity class of a song will be predicted, based on the results of lyrics sentiment analysis. The developed model is a model that has been trained with English songs. Based on the experiment, the model that used the oversampling method achieved an accuracy of 87%, precision of 88%, recall of 87%, and f1-score of 87%.**

***Index Terms*—bidirectional encoder representations from transformers; machine learning; sentiment analysis; song popularity**

I. Introduction

Music streaming platforms have contributed to the increasing public consumption of music, by providing easier access to a wide range of songs. However, the presence of music streaming platforms may not benefit all musicians. According to United Kingdom's market competition observers, competition within the music industry has become tougher because of the music streaming platforms. Only a small part of the musicians affiliated with renowned music labels get advantages from the music streaming system, while the majority of musicians get disadvantages [1]. To address this issue, musicians need to develop new strategies and pay more attention to song's popularity factors, so that their songs appeal to public and generates substantial profits.

Various factors can contribute to a song's popularity, including the lyrics. Lyrics can significantly influence a song's popularity. Songs with specific lyrical patterns are preferred by the community, thus gaining popularity [2]. Deep learning studies have shown that lyric features have a significant impact on song popularity, even more than audio features [3]. Consequently, incorporating lyric features into song popularity prediction enhances accuracy, compared to predictions solely based on audio features [3-4]. The role of lyrics in a song's popularity can be explored through sentiment analysis of the song's lyrics. Sentiment analysis is a computational study that focuses on the sentiment or emotions conveyed in text. Predicting song popularity using sentiment analysis on the lyrics and analysis on the audio features resulted in higher accuracy compared to predictions based solely on audio features. With Naive Bayes, 51% of accuracy can be achieved [5]. In other study which focusing on lyrics, song popularity prediction achieved an accuracy of 73% using BERT [6].

Bidirectional Encoder Representations from Transformers (BERT), a pretrained language model, a deep learning algorithm introduced by Google in 2018, has emerged as a prominent tool in Natural Language Processing (NLP) tasks [7]. BERT has been extensively used in sentiment analysis research. It has outperformed Naive Bayes, Long Short Term Memory (LSTM), and Support Vector Machine (SVM) in sentiment analysis of customer reviews on Amazon [8]. Similar results have been found in product review studies in e-commerce, where BERT outperformed Recurrent Neural Network (RNN), LSTM, and Gated Recurrent Unit (GRU) models [9]. BERT's performance remains highly competitive even against newer NLP models, BERT often surpassing them. For instance, in the analysis of public sentiment regarding the Human Papillomavirus (HPV) vaccine on Twitter, BERT outperformed the Generative Pre-trained Transformer (GPT) [10]. Therefore, BERT is an excellent algorithm for sentiment analysis [11].

This study aims to develop a BERT-based model for predicting song popularity based on sentiment analysis of the lyrics. The study addresses the need of musicians to develop famous songs by considering the lyrical aspect. Songwriters can evaluate their lyrics based on the prediction results, to ensure that their songs resonate well with audiences and have higher chances to become popular. BERT is used in this research since it is superior for sentiment analysis tasks. Additionally, this research served as the following research of a previous study where it was possible to reach an accuracy of 73%, in predicting song's popularity using BERT [6]. This study mainly focuses on English songs. In this study, a dataset which consists of 52,278 English songs was used. The performance of the model is evaluated using accuracy, precision, recall, and f1-score.

II. Literature Review

A. Song Lyrics

Song lyrics are literature in the form of short poetry that conveys the emotions of the songwriter. Lyrics are used as a medium to express the thoughts and feelings of the songwriter to the listeners. Through lyrics, the songwriter can communicate their message to a wide audience [12].

Specifically, song lyrics are short poems composed with specific structures and rules, related to the music and the singer of the song [12]. Song lyrics fall into the category of modern poetry in terms of poetic structure because there are no strict rules regarding rhyme, lines, and stanzas in writing song lyrics. Songwriters have more freedom in writing lyrics, allowing them to develop songs according to their preferences [13].

B. Sentiment Analysis

Sentiment analysis is a computational study that examines the sentiment or emotions expressed in a text. Through sentiment analysis, the polarity of a text can be determined, whether it is positive, negative, or neutral. This polarity value determines the sentiment, emotions, and expressions present in the text, which can be classified as positive, negative, or neutral [14]. Sentiment analysis is also referred to as opinion mining, as it involves analyzing the opinions in a text. Through this process, opinions can be classified as positive or negative [15].

Sentiment analysis can be applied to opinion texts about various topics, products, services, organizations, individuals, and other subjects. By conducting sentiment analysis, valuable insights can be derived regarding individuals' expressions, emotions, and attitudes towards the specific subject matter. Furthermore, the results of sentiment analysis can be transformed into ratings, where opinion texts can be assigned the corresponding rating values [16].

C. Preprocessing

The preprocessing stage is a crucial step in developing Natural Language Processing (NLP) models. It involves data manipulation to become more structured and ready for further analysis [17]. Several tasks are performed during the preprocessing stage, including data cleaning, case folding, stopwords removal, and lemmatization [18 - 19].

1) Data Cleaning: Data cleaning consists of removing numbers, symbols, and irrelevant characters from the text. Non-alphabetic characters are eliminated to prevent interference with the research objectives. Additionally, single characters and excessive whitespaces are removed from the text, resulting in a more organized text structure [17].

2) Case Folding: Case folding involves converting the text into a uniform format, either lowercase or uppercase [19]. This normalization ensures consistency in the text and facilitates easier classification processes.

3) Stopwords Removal: Stopwords refer to commonly occurring words that do not have significant meaning. During the stopwords removal process, these unimportant and commonly occurring words are eliminated from the text. Less relevant words are identified, marked, and subsequently removed, leading to a cleaner text corpus [18].

4) Lemmatization: Lemmatization is the process of transforming a word into its root form [19]. By applying lemmatization, affixes at the beginning, middle, or end of a word can be stripped away, reducing words to their base form.

D. Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model, a deep learning model used in Natural Language Processing (NLP) projects [7]. BERT possesses the capability to handle various downstream tasks, which are specific functions that can generate desired outputs. An example of a downstream task is question answering. To execute downstream tasks, the BERT model requires a sequence of tokenized input. The [CLS] token represents the first token in the sequence, serving as a representation of the entire sequence, while the [SEP] token is used to separate different segments within a sequence. The [CLS] token is a special token that employed at the beginning of the sequence [11].

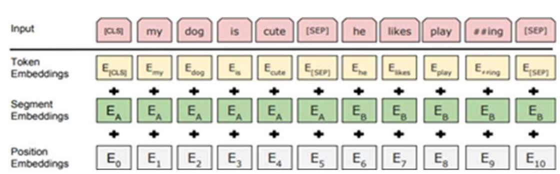

Fig. 1. BERT input representation [11]

The tokenization process involves token embeddings, segment embeddings, and position embeddings. In the token embedding process, the input text is transformed into vectors using WordPiece embedding. With segment embeddings, tokens within the sequence are grouped according to their respective segments. Segments in a sequence can be separated using the [SEP] token. In the position embeddings process, token positions are referenced based on their sequential order within the sequence [7], [11].

BERT modeling consists of two stages: pre-training and fine-tuning. During the pre-training stage, unlabeled data is trained using two unsupervised tasks: masked language model (MLM) and the next sentence prediction (NSP). In the fine-tuning stage, the pre-trained model is further adapted to perform more specific classification tasks (downstream tasks). The constructed BERT model is a multi-layer bidirectional transformer encoder, which consists of 12 layers, 768 hidden states, and 12 self-attention heads [11].

In the pre-training stage, text representations are processed in two different directions, both left-to-right and right-to-left, using the masked language model (MLM) method. The purpose of MLM in BERT is to randomly mask input tokens, replace them with the [MASK] token, then predict the masked tokens [11]. The model predicts the masked tokens by considering the contextual information from both left and right of the masked words. This bidirectional strategy enhances the extraction and the learning process of text features during pre-training [20]. After MLM, the next sentence prediction (NSP) was performed. NSP enables BERT to discover the relationship between one sentence and another [20]. BERT excels at sentiment analysis compared to many other NLP models, by conducting MLM and NSP [11].

In the fine-tuning stage, the pre-trained BERT model is further adapted to more specific tasks (downstream tasks). Specific inputs and outputs are fed into the BERT model, and the model parameters are adjusted to produce the desired output. The [CLS] token representation is specifically fed into the output layer for classification tasks, such as sentiment analysis. Hence, the BERT model can perform sentiment analysis [11].

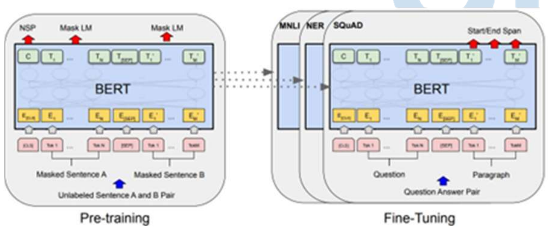

Fig. 2. BERT modeling stages [11]

E. Evaluation Metrics

Evaluation metrics are metrics that can be used to measure sentiment analysis performance [7]. The metric is derived from the analysis of the confusion matrix, which compares the predicted values to the actual values obtained from sentiment analysis model. From the confusion matrix, we can determine the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. These values are used to calculate the accuracy, precision, recall, and f1-score of a sentiment analysis model [21].

Accuracy is an indicator that measures the overall correctness of a model in predicting labels. The calculation is shown in [21, eq. (1)].

$$Accuracy = \frac{TN+}{TP+FP+TN+FN} \tag{1}$$

Precision is an indicator that measures the success of a model in predicting positive elements and how trustworthy a model when it predicts an element as a positive. The calculation is shown in [21, eq. (2)].

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Recall is an indicator that measures the accuracy of the model in predicting positive elements. The calculation is shown in [21, eq. (3)].

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

F1-score is an indicator that measures the overall performance of the model by calculating the harmonic mean of precision and recall. It provides a balanced comparison between precision and recall. The calculation is shown in [21, eq. (4)].

$$F1 - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

II. Methodology

A. Data

This research is using SpotGenTrack dataset, a dataset that consists of songs available on Spotify [22]. This dataset is obtained from Kaggle [23]. Since this study is mainly focused on English songs, only English songs data that were used. By preprocessing the data, 52,278 English songs were found and used for the rest of the study.

B. Data Labeling

Data labeling is a classification process, assigning specific labels to the data. In this research, songs are classified based on their popularity index. Based on the popularity index, songs are classified into less popular, quite popular, and very popular. The songs were classified based on their popularity index, which ranges from 0 to 100. Songs with 0-33 popularity score were classified as less popular songs, 34-66 as quite popular songs, and 67-100 as very popular songs.

C. Preprocessing

Initially, the language of the lyrics is identified using Langdetect, then non-English songs are removed from the dataset. After that, the lyrics undergo several preprocessing steps, such as data cleaning, case folding, stopwords removal, and lemmatization. In the data cleaning process, all non-alphabetic characters are removed from the lyrics. Then, all of the letters in the lyrics are converted to lowercase. Words that have insignificant meaning (stopwords) are also removed from the lyrics using the NLTK. Additionally, the words in the lyrics are lemmatized using the NLTK. Afterward, the lyrics are tokenized to ensure the input requirements of the BERT model.

D. Data Splitting

The dataset is divided into two separate dataset, the training dataset and the testing dataset. The training dataset is used to train the model, while the testing dataset is used to evaluate the performance of the developed model. The data splitting with 80:20 ratio has been commonly used in BERT modeling. This ratio allows the model to achieve optimal performance [24 - 25]. Therefore, the dataset is split into 80% for the training dataset and 20% for the testing dataset in this research.

E. Modeling

Fig. 1 shows the overall of the performed steps in this study. The song lyrics got preprocessed and tokenized to be the BERT-based model's input. In this study, a bert-based-uncased model was used. This model takes the token representations of the input text as input. The output of the BERT model is passed through a softmax classifier layer, which produces song popularity classes as output.

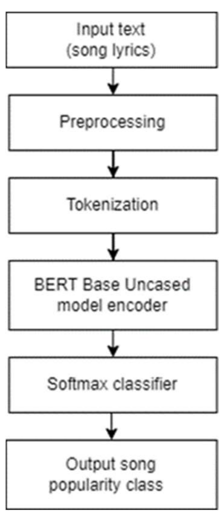

Fig. 3. BERT model architecture

The model's performance then got evaluated using accuracy, precision, recall, and f1-score, as the indicators.

III. RESULTS AND ANALYSIS

A. Implementation

This study is using SpotGenTrack dataset, a dataset that consists of songs available on Spotify from many countries [23]. The songs were classified based on their popularity index. Songs with 0-33 popularity score were classified as less popular songs, 34-66 as quite popular songs, and 67-100 as very popular songs.

After the songs got labeled, the dataset got preprocessed. The language of song lyrics got identified, then the dataset was filtered so it only contains English songs. There are 52.278 English songs data in the dataset. After that, not useful characters were removed from the dataset. The dataset also got case folded, so the lyrics were turned into lowercase. Afterwards, the stopwords from the lyrics got identified and removed. Additionally, the lyrics' form also got identified using Part of Speech (POS) Tagging. After the words got tagged, the words got reduced into their root form.

The lyrics got tokenized using tokenizer, which included a pre-trained BERT model. With using BERT-based-uncased AutoTokenizer which was provided by HuggingFace, we could get the input ids and attention masks. Input ids and attention masks were needed as the input for the BERT model. After passing the preprocess stages, the dataset was split into training dataset and testing dataset. The training dataset was used to train the model. Initially, BERT-based-uncased pre-trained models were imported. The model would receive input ids and attention masks as inputs. There is an additional softmax classifier layer, to produce the output. The output node with the highest probability is then chosen as the predicted label for the input. In this case, the predicted label is the popularity class, whether it's less popular, quite popular, or very popular. This model was designed to process 128 tokens. It's using Adam optimizer, with 3e-5 learning rate. The model's batch size is 32 and trained within 15 epochs.

Models that have been trained were tested using a testing dataset, then got evaluated using accuracy, precision, recall and f1-score as the indicators. Model's performance could also be studied using a confusion matrix. Experiments were conducted to produce model with the best performance. The best model would be implemented in the form of a website application. Fig 4. and Fig 5. are the display of the website application that were built.

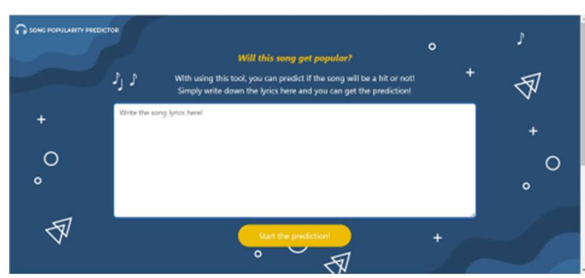

Fig. 4. First part of the application

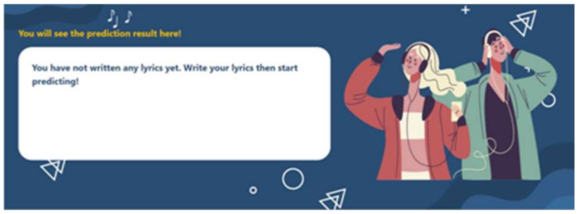

Fig. 5. Second part of the application

B. Experiments

To obtain the best performing model, the modeling process is performed multiple times with different approaches, such as the inclusion or exclusion of stopwords removal and lemmatization. Stopword removal and lemmatization are common steps in NLP modeling and generally improve model performance. However, this may not apply to BERT modeling. Further testing is needed to evaluate the effects of stopwords removal and lemmatization in the context of this study.

Additionally, the data distribution of the dataset can affect the model's performance. If we examine the graph, it can be determined that the dataset is imbalanced. Imbalanced dataset can negatively impact the model's performance. Undersampling and oversampling techniques can be applied to address this issue. The dataset's class distribution is visualized on Fig 6.

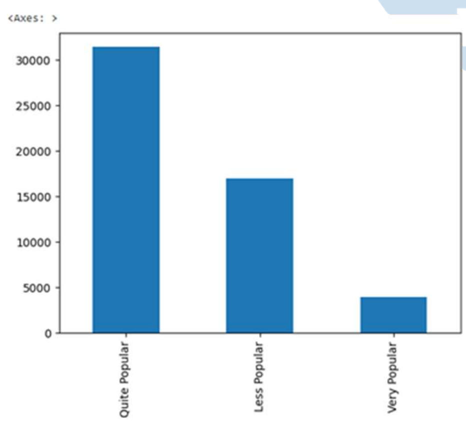

Fig. 6. The dataset's class distribution

Based on the explanations above, a series of experimental scenarios were conducted to model the prediction of song popularity using BERT. The performance of the model was evaluated using accuracy, precision, recall, and f1-score.

1) Without Stopwords Removal and Lemmatization: In the first experiment, the dataset underwent all preprocessing steps except stopwords removal and lemmatization. The model's performance can be evaluated using the confusion matrix. From the confusion matrix, accuracy, precision, recall, and f1-score can be calculated. In this experiment, BERT model achieved accuracy of 84%, precision of 84%, recall of 84%, and f1-score of 83%. The confusion matrix of this experiment can be examined on Fig 7.

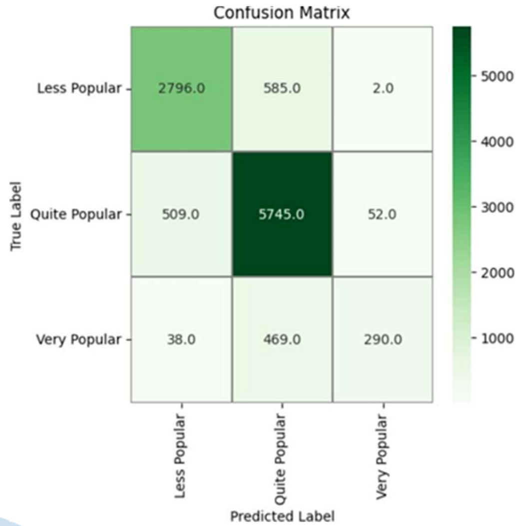

Fig. 7. Confusion matrix for first experiment

2) With Stopwords Removal and Lemmatization: In the second experiment, the dataset underwent all preprocessing steps, including stopwords removal and lemmatization. In this experiment, BERT model achieved accuracy of 84%, precision of 84%, recall of 84%, and f1-score of 84%. The confusion matrix of this experiment can be examined on Fig 8.

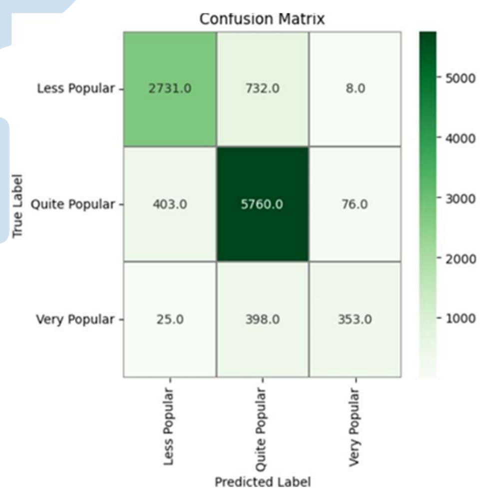

Fig. 8. Confusion matrix for second experiment

If we compare the BERT model's performance with and without stopwords removal and lemmatization, it can be determined that stopwords removal and lemmatization have insignificant impact on BERT model's performance. Both models achieved similar accuracy, precision, and recall values. However, since the model with stopwords removal and lemmatization performs slightly better, stopwords removal and lemmatization were applied in the next experiments.

3) With Stopwords Removal, Lemmatization and Undersampling: In the third experiment, the dataset underwent stopwords removal, lemmatization and undersampling. In this experiment, BERT model achieved accuracy of 85%, precision of 85%, recall of 85%, and f1-score of 85%. The confusion matrix of this experiment can be examined on Fig 9.

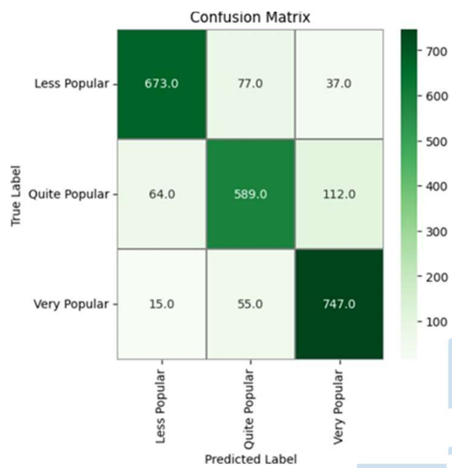

Fig. 9. Confusion matrix for third experiment

When compared to the previous model's performance, it is shown that the model with undersampling performs better. This can be concluded from the higher accuracy, precision, recall, and f1-score achieved by the model with undersampling.

4) With Stopwords Removal, Lemmatization and Oversampling: In the last experiment, the dataset underwent stopwords removal, lemmatization and oversampling. In this experiment, BERT model achieved accuracy of 87%, precision of 88%, recall of 87%, and f1-score of 87%. The confusion matrix of this experiment can be examined on Fig 10.

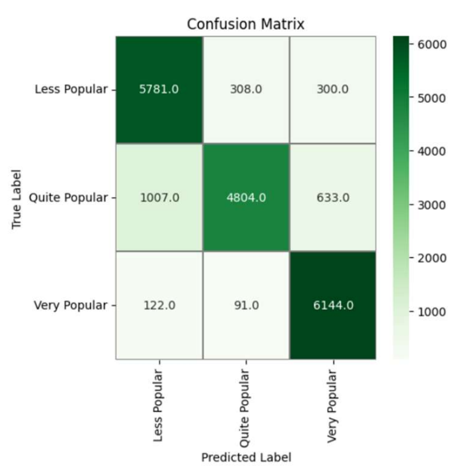

Fig. 10. Confusion matrix for fourth experiment

When compared to model with undersampling, it is shown that model with oversampling achieved higher accuracy, precision, recall, and f1-score. With that, it is concluded that model with oversampling perform better.

C. Evaluation

Based on the experiment results, it can be concluded that the model developed in the fourth scenario, which using stopwords removal, lemmatization and oversampling method, exhibits the best performance. This model achieved an accuracy of 87%, precision of 88%, recall of 87%, and f1-score of 87%. This is the model that will be implemented in the song popularity prediction website application. Thus, the BERT model for predicting song popularity based on sentiment analysis of lyrics has been successfully developed. Furthermore, performance of the different models from the trials can be observed in Table 1.

TABLE I. EVALUATION METRICS VALUE FOR THE EXPERIMENTS

Experiments	Accuracy	Precision	Recall	F1-Score
1	84%	84%	84%	83%
2	84%	84%	84%	84%
3	85%	85%	85%	85%
4	**87%**	**88%**	**87%**	**87%**

III. CONCLUSION

In this study, a model was successfully developed to predict the popularity of songs, based on song lyrics sentiment analysis using Bidirectional Encoder Representations from Transformers (BERT) algorithm. This study used the SpotGenTrack dataset, which contained 52,278 English songs data. By performing preprocessing and oversampling on the training data, the model achieved an accuracy of 87%, precision of 88%, recall of 87%, and an f1-score of 87%. This study outperformed previous study, which only achieved an accuracy of 73% with BERT [6].

IV. FUTURE WORKS

This research could be improved to deliver better results in the future. In this study, a model was successfully developed to predict the popularity of songs, based on English song lyrics sentiment analysis using Bidirectional Encoder Representations from Transformers (BERT) algorithm. In the future, studies about predicting song popularity, which consists of non-English lyrics or lyrics with multiple languages could be conducted. Furthermore, deep learning models could be developed to process a larger number of input tokens. Because of limited resources, this research produced a model that could process only 128 tokens, meanwhile song lyrics could have many more word tokens. Additionally, songs popularity prediction might include other features such as audio features and artists influences, besides of the lyrics. With combining more features, song popularity prediction accuracy might be increased.

References

[1] J. Wilson, "Music industry competition watchdog finds alarming trends across the sector," Forbes. [Online]. Available:https://www.forbes.com/sites/joshwilson/2022/12/05/music- industry-competition-watchdog-finds-alarming-trends-across-the- sector/?sh=4ada9ccc732c

[2] R. A. Nender, P. Rumengan, and G. Latuni, "Struktur musik lagu-lagu koes plus dan pengaruhnya terhadap popularitas dan kelestariannya", Kompetensi: Jurnal Ilmiah Bahasa dan Seni, vol. 1, no. 1, pp. 228–237, 2021.

[3] F. Pachet and P. Roy, "Hit song science is not yet a science." in ISMIR, 2008, pp. 355–360.

[4] A. Singhi and D. G. Brown, "Can song lyrics predict hits," in Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research, 2015, pp. 457–471.

[5] A. H. Raza and K. Nanath, "Predicting a hit song with machine learning: Is there an apriori secret formula?" 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), pp. 112–116, 2020.

[6] I. Sharma, M. Striki, and Z. Gajic, "Hit song classification with audio descriptors and lyrics," 2020.

[7] R. M. R. W. P. K. Atmaja and W. Yustanti, "Analisis Sentimen Customer Review Aplikasi Ruang Guru dengan Metode BERT (Bidirectional Encoder Representations from Transformers)," (Journal of Emerging Information Systems and Business Intelligence), vol. 2, no. 3, pp. 55–62, 6 2021.

[8] G. M P and K. R. Dhanaraj, "Improving the performance of aspect based sentiment analysis using fine-tuned bert base uncased model," International Journal of Intelligent Networks, vol. 2, pp. 64–69, 01 2021.

[9] Y. Liu, , J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e- commerce product reviews by deep learning model of bert-BiGRU- softmax," Mathematical Biosciences and Engineering, vol. 17, no. 6, pp. 7819–7837, 2020.

[10] L. Zhang, H. Fan, C. Peng, G. Rao, and Q. Cong, "Sentiment analysis methods for hpv vaccines related tweets based on transfer learning," Healthcare, vol. 8, no. 3, 2020. [Online]. Available: https://www.mdpi.com/2227-9032/8/3/307

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[12] W. Fachrunissa and N. R. Azzahra, "Analisis makna pada lirik lagu "pelukku untuk pelikmu" karya fiersa besari dalam album pelukku untuk pelikmu," Ta'rim: Jurnal Pendidikan dan Anak Usia Dini, vol. 3, pp. 41–48, 08 2022.

[13] M. Noveanto, H. Sastypratiwi, and H. Muhardi, "Uji akurasi klasifikasi emosi pada lirik lagu bahasa indonesia," Ta'rim: Jurnal Pendidikan dan Anak Usia Dini, vol. 10, pp. 311–318, 07 2022.

[14] A. Z. Amrullah, A. S. Anas, and M. A. J. Hidayat, "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," Jurnal BITe, vol. 2, no. 1, pp. 40–44, 6 2020.

[15] B. Liu, "Sentiment analysis and subjectivity," in Handbook of Natural Language Processing, 2010.

[16] K. Af Ida, "Analisis sentiment terhadap jasa ekspedisi pengiriman barang pos indonesia melalui facebook dengan metode naive bayes dan lexicon based," 2018.

[17] F. Delfariyadi, A. Helen, and S. Yuliawati, "Klasifikasi sentimen judul berita pemberitaan covid-19 tahun 2021 pada media detikhealth," JIEET (Journal of Information Engineering and Educational Technology), vol. 6, no. 2, pp. 50–57, 2022.

[18] M. A. A. Jihad, A. Adiwijaya, and W. Astut, "Analisis sentimen terhadap ulasan film menggunakan algoritma random forest," eProceedings of Engineering, vol. 8, no. 5, 2021.

[19] A. Faadilah, "Analisis sentimen pada ulasan aplikasi tokopedia di google play store menggunakan metode long short term memory," 2020.

[20] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using bert," in Proceedings of the 22nd Nordic Conference on Computational Linguistics. Turku, Finland: Linko¨ping University Electronic Press, Sep.–Oct. 2019, pp. 187–196. [Online]. Available: https://aclanthology.org/W19-6120

[21] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi- class classification: an overview," 2020. [Online]. Available: https://arxiv.org/abs/2008.05756

[22] D. Mart´ın-Gutie´rrez, G. Herna´ndez Pen˜aloza, A. Belmonte-Herna´ndez, and F. A´ lvarez Garc´ıa, "A multimodal end-to-end deep learning architec- ture for music popularity prediction," IEEE Access, vol. 8, pp. 39 361– 39 374, 2020.

[23] D. Mart´ın Gutie´rrez, G. Herna´ndez Pen˜aloza, A. Belmonte Herna´ndez, and F. A´ lvarez, "SpotGenTrack Popularity Dataset," 7 2019. [Online]. Available: https://www.kaggle.com/datasets/saurabhshahane/spotgen-music-dataset

[24] Z. Kilimci, D. Othan, and M. Uysal, "Financial sentiment analysis for predicting direction of stocks using bidirectional encoder representations from transformers (bert) and deep learning models," 12 2019.

[25] C. Oswald, S. E. Simon, and A. Bhattacharya, "Spotspam: Intention analysis–driven sms spam detection using bert embeddings," vol. 16, no. 3, 2022. [Online]. Available: https://doi.org/10.1145/3538491