Bimodal Music Subject Classification via Context-Dependent Language Models

Kahyun Choi[0000−0003−4854−7104]

Indiana University Bloomington

Abstract. This work presents a bimodal music subject classification method that uses two different inputs: lyrics and user interpretations of lyrics. While the subject has been an essential metadata type that the music listeners and providers have wanted to use to categorize their music database, it has been difficult to directly utilize it due to the subjective nature of song lyrics analysis. We advance automatic subject classification technology by employing a context-dependent language model, bidirectional encoder representations from the Transformers (BERT). BERT is a promising solution to reduce the gap between humans and machines' abilities to understand lyrics because it transforms a word into a feature vector by harmonizing the contextual relationship between that word and its surrounding words. The proposed model employs two BERT modules as an ensemble to control the contribution of the two modalities. It shows significant improvement over the existing context-independent models on both the uni and bimodal subject classification benchmarks, suggesting that BERT's context-dependent features can help the machine learning models uncover the poetic nature of song lyrics.

Keywords: Music subject classification · Language models · BERT

1 Introduction

Subject, as a term to represent "what the song is about," has been an important metadata type for music listeners. For example, people have used subject to organize their music library; listeners have used it to search for songs with a particular theme or create playlists under the same subject; and radio DJs have built up stories upon a selected context. However, none of the leading music streaming companies provide the subject metadata of popular songs, while a few websites provide subject information on a small scale (e.g., `songfacts.com`). Given that people are eager to search and browse music based on subject metadata [13], a discrepancy exists between users' needs and services.

The discrepancy originates from a common challenge: the manual annotation of songs is time-consuming and expensive. As for labels that involve more subjectivity, such as mood [11], it becomes harder to collect a large amount of labeled data. Moreover, unlike other tags, no strong relationship exists between subject and audio signals, limiting the annotators and computer algorithms to using only the lyrics.

Another more specific reason that subject is difficult to extract is that song lyrics are often poetic and ambiguous. Indeed, various websites have served millions of users who want to understand song lyrics by reading other users' demystifying postings on stories and meanings of song lyrics (e.g., `songmeaning.com` and `genius.com`). In the past, Choi et al. introduced subject classification systems that utilize both song lyrics and the attentive readers' comments collected from those websites [3][4][5]. In those works, users' comments turned out to be more useful than the song lyrics themselves when used as the input for an automatic subject classification system.

While the previous research introduced the potential of user-generated data for automatic subject classification, the classification system needs improvement because the methodology is missing important aspects of learning from the text as a sequence. Thus, we propose to use bidirectional encoder representations from the Transformers (BERT) [8], which are widely known to provide a context-dependent language model from word sequences via the self-attention mechanism [20]. We also address the bimodal use case by introducing a trainable blending parameter to combine the two classification results from both modalities as an ensemble: lyrics and interpretation. To this end, we focus on the following research questions:

- Q1. Does the BERT-based softmax classifier outperform the SVM model using TF-IDF features?
- Q2. Is the weighted ensemble method effective for combining bimodal classifiers?

2   Related Work

From the mid-2000s, a few papers have proposed automatic music subject classification systems. In 2005, Mahedero et al. proposed the first subject classification system that relies solely on song lyrics [14]. This seminal work introduced music subject classification as a challenging problem and showed that basic machine learning models and small datasets are not enough to build a robust classification system. There have been some unsupervised topic analysis methods as well, such as using nonnegative matrix factorization [12] or latent Dirichlet allocation [6]. Although more scalable, these methods are limited to the unsupervised setup. More recently, Choi et al. discovered that user-generated interpretation data is more useful than song lyrics [4] and proposed a bimodal system that benefits from both lyrics and interpretations [5]. This kind of approach showed the potential of using people's interpretation of song lyrics as an alternative source of subject-related information. The bimodal system verified that interpretations are more useful than song lyrics, while the combination of the two only slightly improves the classification accuracy. Fig. 1 summarizes the SVM- and TF-IDF-based bimodal subject classification.

However, the subject classification systems discussed so far did not consider the order of the words within the sequence. For example, the TF-IDF (term frequency-inverse document frequency) feature does not preserve the sequential
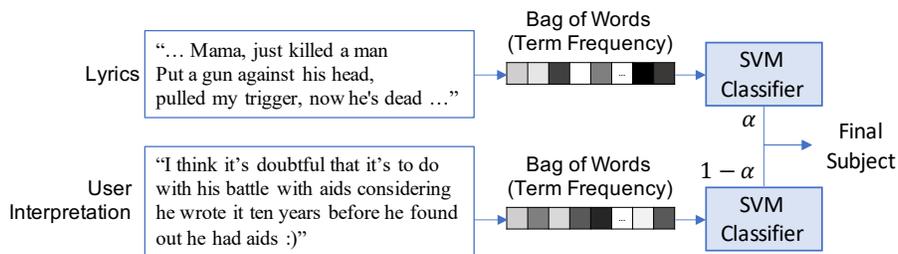
Fig. 1. The bimodal classification framework using two SVM classifiers on the two input sources: lyrics and their interpretations in [5]. The harmonization of the two classification results is done by performing a manual search for the optimal late-fusion parameter α via a ten-fold cross-validation process. The lyrics excerpt is from the song, "Bohemian Rhapsody" by Queen. In comparison, one of the top comments in `songmeanings.com` is also presented, which reads between the lines.

order of the words. Furthermore, support vector machines (SVM), as the main classification method, worked only as a non-sequential classifier [19]. To overcome the limitations, a recurrent neural network, the long short-term memory (LSTM) [10], was employed to learn from the sequence of word embedding vectors [3]. While the fastText-based word embedding vectors should, in theory, capture the semantic relationship between words [2], and the LSTM model is powerful for learning the sequential information, this new model was not comparable to the previous non-sequential models. It is mainly because of the difficulty in training LSTMs using a limited amount of labeled data.

The previous work's limitations led us to employ BERT as a more powerful pretrained feature extraction method. The benefit of using BERT in our work is threefold. First, BERT is an extensive and powerful deep neural network model, pretrained from large text corpora to generalize well to unseen problems with little adjustment. It means that our BERT-based classifier will not suffer much from overfitting, while it is expected to perform better than the previous methods. Second, as a Transformer-based self-attention model, it learns the context from the word sequence, minimizing the classifier's role. In other words, there is no need to train and use LSTM or SVM classifiers. Third, as a neural network framework, it is straightforward in extracting two sets of features from the two modalities using two BERT models and then finding the optimal combination ratio between them.

3   The Proposed Method

3.1   Dataset

We follow the data preparation process proposed in the latest work that handled subject classification on bimodal data [5]. We used two types of input data: song

lyrics and user comments provided by `LyricFind`[1] and `songmeanings.com`[2], respectively. `Songmeanings.com` has served millions of music lovers who want to discuss the meanings of song lyrics. Users can post their interpretations of song lyrics in the comment section for each song. They can also rate other users' comments so that the most highly-rated comments would appear on top. Some songs have many comments, while some do not. We collect up to the top ten comments from each song and then concatenate them as the BERT models' input sequence.

For the subject labels, we refer to the music database on `songfacts.com`. `Songfacts.com` provides a searchable database of songs curated by experts. "About" is one of their browsing options, which corresponds to the song lyrics' subject. Among the 206 "about" categories, we selected the eight most popular subject categories with more than 100 songs: {`Religion`, `Sex`, `Drugs`, `Parents`, `War`, `Places`, `Ex-lover`, and `Death`}. The number of songs per class is limited to 100 to prevent our classifiers from favoring more populous categories.

3.2   Classification Setup

BERT BERT [8] is one of the latest natural language processing (NLP) models that use the encoder part of the Transformer model [20] for language modeling. BERT showed state-of-the-art results in various NLP tasks, such as document classification [1], by overcoming the limitations of its predecessors, including the TF-IDF and context-independent word embedding methods. Conceptually, it is similar to the other word embedding techniques, such as Word2Vec [16] and fastText [2], as it can learn the semantic relationships between words. However, unlike context-independent models that learn an embedding vector, which aggregates all the meanings that a word is associated with, BERT learns the embedding vector based on the word's context within a sentence.

$$\mathbf{v}_i = F(w_i) \tag{1}$$

$$[\text{CLS}, \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_i, \cdots, \mathbf{v}_N] = \text{BERT}([w_1, w_2, \cdots, w_i, \cdots, w_N]). \tag{2}$$

In (1), for example, a context-independent model learns the mapping function $F(\cdot)$ that converts the i-th word w_i in a sentence into an embedding $\mathbf{v}_i$. In this process, the model does not consider the context that w_i belongs to, which can significantly change the word's meaning. On the other hand, as shown in (2) the BERT model is a function of the entire word sequence so that the embedding vectors represent the relationships among the words within the same sentence. To this end, BERT always takes the entire sentence as input instead of an individual word. Note that BERT also predicts a sentence-specific vector CLS, which works as a summary of the input sentence. We use it as the final feature vector for the

softmax classifier. Likewise, as we do not need all the embedding vectors, except for the CLS embedding, our simplified BERT function is defined as follow for the rest of the paper:

$$CLS = BERT(\mathbf{w}), \tag{3}$$

where $\mathbf{w} = [w_1, w_2, \cdots, w_N]$.

Another important advantage to note is that BERT employs the Transformer model that uses the powerful self-attention mechanism. Compared to its LSTM-based context-dependent predecessor, i.e., embeddings from language models (ELMo) [17], the self-attention mechanism exhibited improved performance. As BERT models are pretrained on massive datasets and are publicly available, NLP tasks with a small dataset can benefit directly from BERT even without further finetuning.
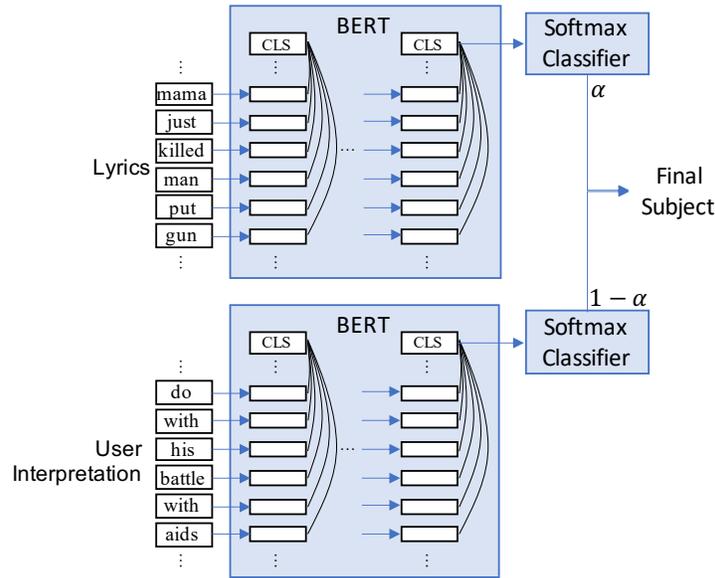
Fig. 2. The proposed bimodal classification framework using two BERT classifiers on the lyrics and interpretation input streams. The fusion of the modalities is performed as a part of the training process by defining the ensemble weight α as a trainable parameter.

The Ensemble Method We combine two classification results from two BERT models: the lyrics and interpretation modalities, $\mathbf{w}_L$ and $\mathbf{w}_I$, respectively. Hence, the final classification result is a weighted average of the two classification results,

$$\alpha \, \mathrm{softmax}\left(BERT(\mathbf{w}_L)\right) + (1 - \alpha) \, \mathrm{softmax}\left(BERT(\mathbf{w}_I)\right), \tag{4}$$

where $0 \leq \alpha \leq 1$ is the ensemble weight that defines the contribution of the two modalities. We define it as a trainable parameter with an optimal value found through the model training process, along with the softmax classifier's parameters. Finally, the ensemble results in a probability vector over all eight classes. Its largest element is associated with the predicted class the example belongs to. Fig. 2 describes the proposed ensemble model using two BERT modules to handle lyrics and interpretation.

The Experimental Setup We employed the pretrained BERT model available as a part of the `ktrain` package [15]. Due to the heavy computational requirement, we did not attempt to finetune the BERT model for our classification problem. Instead, we modified the original unimodal BERT classifier into a bimodal system that eventually runs the pretrained BERT module twice. The two CLS embedding vectors are then fed to the two softmax classifiers to predict subject classes, which are eventually combined as an ensemble. The one cycle policy was used for optimization [18]. For a fair comparison with previous works, we followed the same ten-fold cross-validation process. All implementations are based on the `Keras` deep learning framework [7].

4   Experimental Results and Discussion

Table 1 summarizes the performance of the proposed BERT-based models compared against the previous SVM-based model on the music subject classification benchmark used in [5].

Table 1. The comparison of the proposed BERT-based model and the previous SVM-based model. The SVM-based model's performance was reported in [5]

| System	The BERT-Based Model			The SVM-Based Model [5]		
Input	Lyrics	Interp.	Bimodal (Trainable Ensemble)	Lyrics	Interp.	Bimodal (Late Fusion)
Death	44	57	60	29	51	50
Drugs	50	71	78	36	69	70
Exlovers	52	68	70	36	67	68
Parents	37	65	63	34	57	60
Places	57	62	66	49	58	61
Religion	52	78	78	35	70	70
Sex	67	66	80	65	70	73
War	73	77	79	65	76	79
Average	54	68	71.8	43.6	64.8	66.4

Overall Performance Comparison The proposed BERT-based classifier improves the overall performance across all input types. It prefers the interpretations over the lyrics, which is the same preference reported in the SVM model on the TFIDF features. Notably, the proposed method extracts subject information from lyrics more effectively (54%) than the traditional method (43.6%). As for the interpretation-only input, the BERT classifier increased the accuracy from 64.8% to 68%, which is not as significant as the lyrics-only case. Since we use a pretrained BERT model for both types of input, the more considerable improvement on the lyrics-only input does not necessarily mean that BERT works better on lyrics than interpretations. Indeed, the classification accuracy of the interpretation-only model is still much higher than the lyrics-only model. We believe that this drastic improvement in the lyrics model comes from the potential causes, such as a) lyrics are challenging to analyze, b) there is more room for improvement, and c) BERT may have successfully interpreted the latent meanings of some poetic words in lyrics using the contextual information. These findings suggest that BERT might be a promising method for extracting latent information from other texts that are difficult to analyze due to their brevity and complexity, such as poetry.

Bimodal Classification When both text sources are used, the proposed BERT-based bimodal classifier's accuracy reached 71.8%, which is 5.4% higher than the competing bimodal SVM classifier's performance. The proposed trainable ensemble is efficient since it estimates the ensemble weights as a part of of the neural network optimization. On the other hand, the previous late fusion technique tediously examines all possible mixing ratios. The estimated ensemble weight turned out not too different from a simple average: $\alpha = 0.498$, whose standard deviation is 0.003. Therefore, we believe that our bimodal classification system directly benefits from the significant improvement in the lyrics modality rather than the fusion mechanism itself.

Confusion Analysis The proposed BERT-based classifier shows robust improvement in most of the subject categories across all three input types, except for two cases: the interpretation-only case for the `Sex` category and the bimodal model for the `War` category.

In particular, we found that both the previous and proposed classifiers shared the same sets of difficult and easy subjects. For both classifiers, the easy subjects were `War`, `Sex`, `Drugs`, and `Religion`, while the difficult ones were `Death`, `Parents`, `Places`, and `Ex-lovers`. `War` was the easiest subject in the previous research with 79% accuracy, the same as the proposed research. As for `Sex`, `Drugs`, and `Religion`, their accuracy was around 70%, but now, it reached almost 80%. The most difficult subject was `Death`. Its accuracy was only 50%, but with the proposed classifier, it was 60%. Both classifiers have the same difficult and easy sets of subjects, indicating that songs with difficult subjects may be associated with multiple subjects. For example, if a song in the `Death` category is also talking about `Religion`, it is more realistic to assign multiple labels. This suggests

that a multi-label classifier might be a more sensible choice than a single label classifier for the subject classification problem.
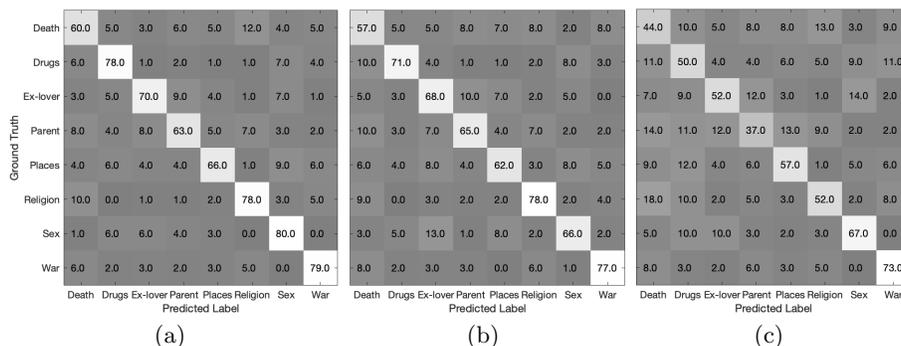
Fig. 3. Confusion matrices from the proposed classification systems (a) the bimodal system using trainable ensemble weight α (b) the interpretation-only unimodal system's results (c) the lyrics-only case.

We also examined the confusion matrices among subject categories to determine which pair of categories is most confusing to our proposed classifiers (Fig. 4). The confusion matrices showed that the most confusing pair of categories was `Death` and `Religion`, and they were consistently misclassified by each other across all three different input types. The fact that `Death` and `Religion` have a close relationship in human history [9] might have led to their possible coexistence in many songs. As for the unimodal cases, `Death` was often misclassified as `War`, `Drugs`, or `Parents`, and vice versa. However, the bimodal classifier reduced such confusion to some degree, which indicates that the bimodal classifier effectively benefits from the complementary relationship between the two unimodal classifiers.

5   Conclusion and Future Work

In this work, we found that BERT is powerful in extracting subject information from song lyrics, which has been known to be difficult to understand for humans and machines. Our proposed method showed greater classification accuracy overall, but more saliently on the lyrics modality, where the traditional methods left more room for improvement. We also proposed an efficient ensemble method, which showed reasonable improvement over both unimodal systems. While the pretrained and fixed BERT features have not been trained or finetuned from our dataset, they significantly improved the previous classification models that were customized to the dataset. This indicates that further finetuning of BERT could improve its performance. As future work, we will conduct a more detailed

analysis by using representative features of subject categories to identify the contribution of words. We will also explore a bigger benchmark dataset built from other websites, such as `genius.com`. Finally, an expansion to the multi-label classification setup is another promising direction.

References

1. Adhikari, A., Ram, A., Tang, R., Lin, J.: Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398 (2019)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
3. Choi, K., Downie, J.S.: Exploratory investigation of word embedding in song lyric topic classification: promising preliminary results. In: Proceedings of the IEEE/ACM Joint Conference in Digital Libraries (JCDL) (2018)
4. Choi, K., Lee, J.H., Downie, J.S.: What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. In: Proceedings of the IEEE/ACM Joint Conference in Digital Libraries (JCDL) (2014)
5. Choi, K., Lee, J.H., Hu, X., Downie, J.S.: Music subject classification based on lyrics and user interpretations. In: Proceedings of the American Society for Information Science and Technology (ASIS&T) (2016)
6. Choi, K., Lee, J.H., Willis, C., Downie, J.S.: Topic modeling users' interpretations of songs to inform subject access in music digital libraries. In: Proceedings of the IEEE/ACM Joint Conference in Digital Libraries (JCDL) (2015)
7. Chollet, F., et al.: Keras. https://keras.io (2015)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: The annual conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2019)
9. Garces-Foley, K.: Death and religion in a changing world. Routledge (2014)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory **9**(8), 1735–1780 (1997)
11. Hu, X., Downie, J.S.: Exploring mood metadata: Relationships with genre, artist and usage metadata. In: ISMIR. pp. 67–72. Citeseer (2007)
12. Kleedorfer, F., Knees, P., Pohle, T.: Oh oh oh whoah! towards automatic topic detection in song lyrics. Proceedings of the 9th International Conference on Music Information Retrieval pp. 287–292 (2008)
13. Lee, J.H., Downie, J.S.: Survey of music information needs, uses, and seeking behaviours: preliminary findings. In: ISMIR. vol. 2004, p. 5th. Citeseer (2004)
14. Mahedero, J.P., MartÍnez, Á., Cano, P., Koppenberger, M., Gouyon, F.: Natural language processing of lyrics. In: Proceedings of the 13th Annual ACM International Conference on Multimedia. pp. 475–478 (2005)
15. Maiya, A.S.: ktrain: A low-code library for augmented machine learning. arXiv preprint arXiv:2004.10703 (2020)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
17. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proc. of NAACL (2018)

18. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820 (2018)
19. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural processing letters **9**(3), 293–300 (1999)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)