Sentiment of Song Lyrics: Context-Based Model vs Bag-of-Words Model

Hil Alcee
alcee@berkeley.edu
Nov 2022

Abstract

Music from Top 100 Billboard songs have been the soundtrack and reflection of societies feelings and mood overtime. However, recent studies have indicated that lyrics from these popular Billboard songs have become more negative over time. These sentiment analyses were conducted using a bag-of-words approach. To better score the sentiment of each lyric, the researchers utilized the valence on single word text by leveraging the Linguistic Inquiry and Word Count (LIWC). Applying sentiment analysis based on the aggregation of a collection of words in a lyric could potentially lose the context of the overall meaning of the lyric. Would the sentiment of lyrics from Top 100 Billboard from 1965 to 2015 still reflect a similar probability distribution if we used a context-based neural network model and transformer model rather than a bag-of-words model? In this paper, we will construct a LSTM neural net model and fine-tune a Bert Model with and without a layer of CNN to predict the sentiment of each song lyrics from the Top 100 Billboard songs from 1965 to 2015. Finally, we will collect the sentiment distribution from the deep learning models and compare the distribution to the LIWC model to answer our hypothesis question above.

1 Introduction

NLP Deep Learning has made alot of progress in the last few years, and gives an opportunity to better mine the context in language. LSTM, and most recently BERT models, have been successful in classifying sentiment for product reviews, movie reviews, and even tweets. This also provides a great opportunity to attempt to classify the sentiment in music lyrics as well. In this paper, we will leverage several methods in classifying the sentiment in the lyrics from the Top 100 Billboard songs from 1965 and 2015 using context based modeling. To fine-tune our model, we will utilize the MuSe dataset and lyrics from Genius.com. Next, we'll work with a LSTM model and a BERT model (with and without a CNN layer) to classify our sentiments on each lyric. Lastly, we'll collect the probabilities of the lyrics of each song and compare both visually and through paired statistical t-test.

2 Background

Music lyrics from english speaking songs have been a strong reflection of the culture and population over the year. The sentiment from the lyrics have also provided insight if the listeners are drawn more to positive lyrics or negative lyrics over a course of time. The study "Cultural Evolution of Emotional Expression in 50 years of Song Lyrics" **1.(Brand, Acerbi, Mesoudi - 2019)** indicates that song lyrics of Top 100 Music Billboard songs from 1965 to 2015 have become more negative over time. However, the researchers approached this study by leveraging a bag-of-words approach with the billboard songs. When conducting this study, the researchers leveraged Linguistic Inquiry and Word Count (LIWC) to evaluate the sentiment of lyrics of a song. They consider the valence of the lyrics as a measure of sentiment for the lyrics. The analysis scores the valence of each word and whether it is negative or a positive valence. Next the researchers' models are aggregated binomial models. The words are aggregated within each song as each word of the song are modeled as the binomial probability of being positive (or not). The negative models model the likelihood that each word in a song is negative (or not). This factors in the fact that each song has a different number of words, and negates any need for averaging over words and songs. This helps song lyrics from being overstated and understated depending on the amount of words in the lyrics. Each of their models have a similar combination where the probability that any given word in a song is positive (or negative) can be predicted by the average number of positive (or negative) words of previous top songs or artists in the preceding three years of the billboard list. However, only averaging the valence of the words may lose context. For example, Olivia Rodrigo's 2021 top single "Good 4 u" chorus shows a positive tone score of 12.12 and a negative tone score of 1.52 because it includes words like "Happy", "Good", "Cared", and "Healthy". These scores do not properly reflect that in the chorus Rodrigo is showcasing passive aggressive tone and sharing in the same chorus "crying on the bathroom floor" or "I've lost my mind". In contrast, Kendrick Lamar's "Alright" chorus ,which have lyrics about overcoming adversity, has a 0 of LIWC positivity and a 4.39 of negative tone. Based on these outcomes, we can see that this is a challenging task.

Other researchers have attempted a similar study with a similar bag of words of approach. Using Diction 7.0 on lyrics during Covid 19 shutdown during 2020, researchers have found lyrics had significantly more negative valence **2.(Putter, Krause, North - 2022)**. These are definitely acceptable approaches, however we have more context-based technology available to us to try as alternatives.

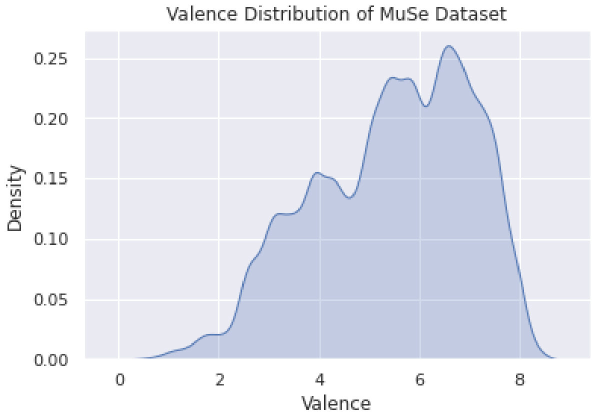
3 Method

3.1 Data

For this study, we will fine-tune our models on the datasets "The Musical Sentiment Dataset" (MuSe) **3.(Akiki, Burghardt - 2021)** and associated song lyrics imported from API of Genius.com. There is also additional song data from Spotify API for some analysis around the year of the songs. Following fine-tuning, we will predict on the lyrics from the same Top 100 Billboards song lyrics from 1965 to 2015. For probability distribution comparisons, we will run the Billboard lyrics through the latest version of LIWC: LIWC 2022.

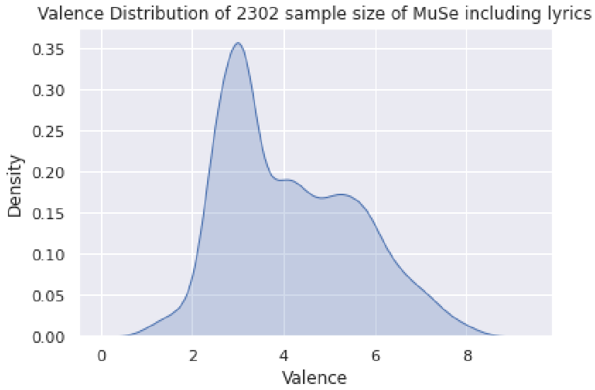
The MuSe dataset contains sentiment information for 90,001 songs. After joining with the Spotify dataset, we are able to classify the years of the songs in the the MuSe dataset. We are able to confirm that the MuSe dataset contains songs with release dates from 1920 to the 2020s. The researchers computed scores for the dimensions of valence (the pleasantness of a stimulus), dominance (the degree of control exerted by a stimulus), and arousal (the intensity of emotion provoked by a stimulus) as defined by the research paper "Norms of valence, arousal, and dominance for 13,915 English lemmas" **4.(Warriner, Kuperman, Brysbaert - 2013)**. The tags for these dimensions were generated by users that are available for each song via Last.FM. The aforementioned researchers were confident in these Last.FM user tags related to valence thanks to the in-depth research found in paper "Music Mood Dataset Creation Based on Last FM Tags". **5.(Cano, Morisio - 2017)**. For our target sentiment, we will be using valence as what was used in the paper **1.(Brand, Acerbi, Mesoudi - 2019)**. In this approach, we agree with the researchers in leveraging valence as a measurement of sentiment. As stated above, valence is the pleasantness of a stimulus (english lemmas in this context) and users tagging songs from Last.FM reflected that when listening to the lyrics. The valence in the MuSe dataset scales from 0 to 10, with a mean of 5.45 and a max and min of 8.47 and .24 respectively. We will make the assumption that anything greater than or equal to 5, is a positive valence. Anything less than 5 will be considered a negative valence. With these assumptions and this 90k plus collection of songs, we can see that the valence distribution skews toward a more positive valence Figure 1.

Figure 1: Valence Distribution of MuSe Dataset



Next, we leverage the Genius.com API to import lyrics associated with the MuSe dataset. The MuSe dataset is a quite massive collection of songs and becomes too much for Genius.com API to allow. For the sake of this study, we have a sample of 2302 songs. This obviously can impact our distribution and our results. We can notice the impact of the distribution from the main dataset to this smaller sample size in Figure 2. It skews toward a more negative valence.

Figure 2: Valence Distribution of 2302 sample size of MuSe including lyrics



However, either distribution is not a normal distribution. This will force us to call on the theorem of Chebyshev's inequality and have confidence that at least 75 pct. of all the data in the smaller sample falls at least 2 standard deviations from the mean. It should be enough data to help represent the population of context in lyrics for our training set.

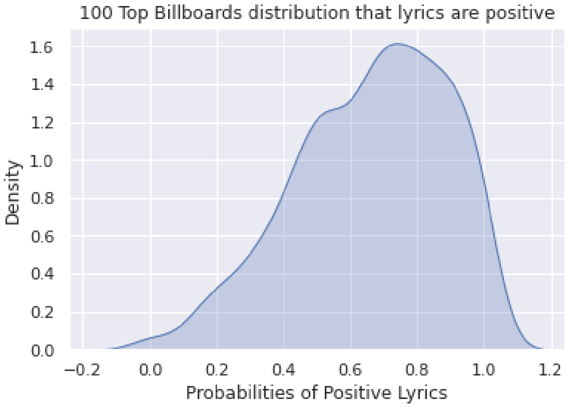
Finally, to compare our models to the researcher's LIWC model, we leverage the Top 100 Billboard 1965 to 2015 dataset. The Billboard dataset was curated by scraping the Billboard's top songs from Wikipedia and coalescing associated lyrics from websites like metro lyrics, song lyrics, and lyrics mode. To extract the probability of a positive lyric, the positive valence feature ("tone pos") and the negative valence feature ("tone neg") were leveraged. The probability calculated for each song:

$$LIWCProb = tonepos/(tonepos + toneneg)$$

3.2 Data Pre-Processing

The lyrics from the training set were further processed by removing punctuations, back slashes (as lyrics have stanzas), and lower-casing all words. The lyrics in the Billboard dataset were also processed in the same way as the training set by removing punctuations and lower-casing all words. Following these normalizations, we leverage the Bert Tokenizer to convert our lyrics into token input ids to be recognized by the deep learning models. Afterwards, utilized a function "select min length examples" to truncate the vectors to the tokens the tokenizer recognizes as the relevant token ids. This is done by filtering by the underlying attention masks of the token values. For consistency, the same was applied to the billboard dataset. After drawing out the relevant values, the dataset shrunk from 5053 songs to 916 songs. This gives us a more concentrated dataset to use for our models, to predict, and to compare our distributions. Figure 3 depicts that even with the lower number of songs, the distribution still aligns with the original MuSe dataset.

Figure 3: 100 Top Billboards probability distribution after filtering to relevant tokens



Next, after tokenization and extracting the input ids, the data was split into a 70-30 train-development dataset split with a random state of 42 to reproduce results. Following the split, the development data was further split 50-50 into a development-testing split. We will be gauging the validation accuracy of our models using the development split, and test accuracy on the testing set.

3.3 LSTM and BERT Models

In the following sub-sections, we will discuss utilizing a LSTM model, base BERT pooled output model, and a BERT model with a CNN layer. The sub-sections will talk about the reason behind the choice models, brief construction of the layers, hyper parameters and accuracy measurements.

3.3.1 LSTM

To begin our modeling process, we are starting with an LSTM as our baseline model. Lyrics are essentially a sequence of words that build upon each other to establish meaning. LSTMs are a strong candidate in its ability to carry forward the meaning of a sequence while capturing new information as it moves forward. After gathering much of the context of the sequence, hopefully we will be able to derive the actual context of the lyrics.

The LSTM model consist of three hidden layers, with three dropout layers following each of the three LSTM hidden layers with a rate of .3. This will help the model to mitigate overfitting. The idea of having more hidden layers is to reduce bias in our model and better fit in relation with our training set to our target sentiments. This will be important when we leverage the model to predict on the unseen Billboard lyrics data. For our baseline, especially since we are hoping this combination will balance between the bias and variance trade off. Our final output layer consist of a sigmoid activation layer as our target variables are a binary output. Finally, we compile the model utilizing a common deep learning choice of an "Adam" optimizer with the learning rate of .00005. The loss optimizer is using a binary cross entropy. We are running our model with an epoch of 8 with batch sizes of 32 to better iterate over our training set.

3.3.2 BERT Pooled Based Model

Next, we are leveraging the pre-trained BERT model and using our training set to fine tune the model for our respective target variables.

We are leveraging the BERT base uncased model for this study in comparison to the BERT large uncased model to keep it within the confines of our compute power. The uncased model was selected as our lyrics are uncased after the preprocessing. The model is utilizing the the pooled token output rather than the CLS token output. When testing both, the pooled token produces a better validation accuracy. To keep our study concise, we will focus on the pooled token output for our BERT Base model.

Following the output, our model consists of one hidden layer. Our hidden layer consists of a Relu activation layer followed by a dropout layer of .4 to reduce overfitting. The final layer is using a sigmoid activation as our target variable is a binary output. To better fine tune this model, this hidden layer was include with respective to dropout to better balance the bias-variance trade off. This is especially true since we have a smaller sample size to fine-tune on. Finally, we are compiling the model with the Adam optimizer with a learning rate of 0.0005. The learning rate is slower for this model as to not overshoot our optimization since we are only using one layer. The loss optimizer for this model is using a binary cross entropy. The loss optimizer is using a binary cross entropy. Like the LSTM model, we are running our model with an epoch of 8 with batch sizes of 32.

3.3.3 BERT Model with CNN layer

Finally, we are leveraging the pre-trained BERT model and adding a CNN layer to this model. The idea is that we can leverage the CNN layer to really focus in and provide better feature selection from our BERT model outputs to provide to our hidden layers. For this model, we will utilize the CLS token vector so the CNN layer can accept matrix shape and process through the convolutional layer. It will process through its own pooling layer to work as features for our hidden Relu layer.

Next, we are passing in a list of window sizes and number of windows for the CNN layer to iterate over and slide across the inputs to extract important features. Its iterating over a combination of 100, 100, 50, 25 filters and window sizes of 3, 5, 10, and 20 respectively. The idea is to start with a large number of smaller windows to extract information of the vectors at a micro scale, and later use the smaller number of large windows to cover the macro layer of context. Each iteration pass through a layer with a Relu activation, extracting information into a pooling layer, and then concatenated into a final convolutional output.

After the CNN layer, we pass our convolutional output through a hidden layer with a Relu activation followed by a dropout rate of .4. The final output layer utilizes a sigmoid activation function for a binary target variable. Finally we compile the model with an Adam optimizer with a learning rate of .0005.

4 Results

After running the respective models, we measure the accuracy on the model's ability to predict the sentiment from the test data. We are using these traditional measures to estimate the strength in our models. If our models are strong enough, we can have more confidence in their respective probability distributions. Afterwards, we will use the model with the highest accuracy to predict on the Billboard data to extract the the probabilities. This will allow us to compare the probability distribution from the context-base model applied to the Billboard lyrics compared to the probability distribution from the LIWC model. Based on this analysis, we will gauge if the distributions are similar. And to check our hypothesis, we will conduct our respective paired t-test.

Results of our models are presented in Table 1.The Bert Model with CNN Layer outperforms the other two models. The LSTM model alone performed quite poorly. This is mostly due to the model only able to maintain the context of the lyrics in its short term memory as it moves forward through the sequence. We have more success with the Bert models with their self-attention querying capabilities. They are able to maintain the context of the sequence before and after the token evaluated. The two Bert models performed much better and are very close in their accuracy. It is worth noting that the CNN Bert model performs the best in precision accuracy, but face a bit of challenge in minimizing false negatives in its recall. Where the biggest challenge for all of the models are the AUC scores. The models are having a difficult time clearly classifying the target values. When evaluating the incorrect predictions in the test data and reviewing the lyrics, some songs may be mislabeled by the users themselves that the model is actually predicting correctly. For example, a song that is consistently showing as incorrectly labeled is ”Rain Water” by Brother Ali. Its a song about how an individual is losing hope. The model is responding to the negative terms in the lyrics, however the users have tagged it slightly above the valence threshold of 5 (score of 5.02) resulting in a positive tag. 5.02 still may be a little high for such a dire set of lyrics. ”Love is Blind” by Eve is another example with the same score, and potentially even more negative lyrics. For a future study, there may be opportunity to have a lyrics dataset tagged by lyricists rather than users. However with reference to the previously mentioned Olivia Rodrigo’s and Kendrick Lamar’s songs in the introduction section, the Bert CNN model predicted the expected sentiments. Rodrigo’s song ”Good 4 You” has an underlying negative connotation, and the model was able to predict it successfully with almost a 99 pct probability. Kendrick Lamar’s clean censored version of song ”Alright”, was predicted accurately as a positive sentiment. The model predicted ”Alright” with 53 pct probability as the model was still able to capture negative connotation weaved within the lyrics. With this all being said, the model could potentially improve being trained or fine-tuned on a larger amount of training data. We will touch upon this more in the conclusion section.

For this study, we will utilize the probabilities from the model with the highest accuracy to test our hypothesis.

Table 1: Accuracy scores of context-based models predicting on test data

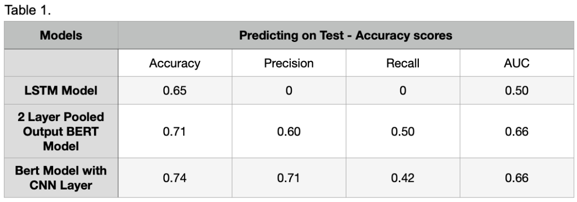
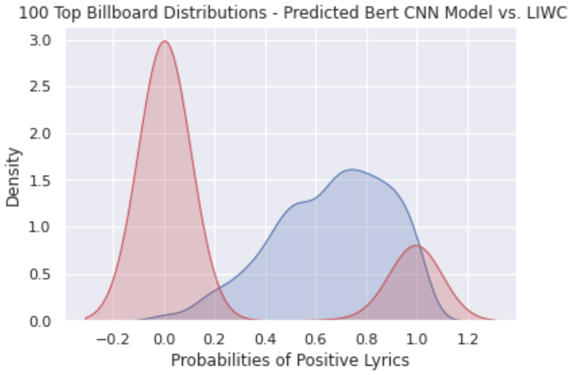


Table 1.

Models	Predicting on Test - Accuracy scores			
	Accuracy	Precision	Recall	AUC
LSTM Model	0.65	0	0	0.50
2 Layer Pooled Output BERT Model	0.71	0.60	0.50	0.66
Bert Model with CNN Layer	0.74	0.71	0.42	0.66

Figure 4: 100 Top Billboard Probability Distributions - Predicted Bert CNN Model (in red) vs LIWC Model(in blue)



As Figure 4 depicts, the distributions between the BERT CNN Model vs the LIWC are different. In fact, the BERT CNN Model is showing there is a greater distribution of less positive lyrics. To confirm what Figure 4 is depicting, we will run a paired t-test between the distributions to test our hypothesis.

For this study, a paired t-test will be conducted as we are using the same set of Billboard lyrics, but comparing the probability results from the two different models: Bert CNN Model and LIWC model. The paired test will be conducted as:

H0:

$$\mu_1 = \mu_2$$

”The two population means are equal”

H1: two-tailed:

$$\mu_1 \neq \mu_2$$

”The two population means are not equal”

After conducting the paired t-test, and with a p-value much lower than .05 (p-value = 3.51e-151), we can conclude that we reject the null hypothesis. This indicates that the two distributions do not have identical mean values. In other words, the probability distribution between the context-based models and the LIWC model do not share a similar probability distribution of the Top 100 Billboard lyrics from 1965 to 2015. This may be indication that the context-based model can potentially derive more of a context than what the LIWC based models are providing.

Lastly, the overall theme of the study in which this research was inspired **1.(Brand, Acerbi, Mesoudi - 2019)** depicts a rising trend in negative lyrics. The trend is not the focus of the paper, however it is worth noting in Figure 4 that the probability distribution of context-based model skews mostly to the left. This could indicate that the overall distribution of lyrics in the Billboard dataset may be predicting as less positive (or negative sentiment). There is also a smaller group in the positive leaning distribution as well. We do not know where these probabilities occur in the trend timeline, especially using the researches probability model, but songs could potentially be more clearly defined as more negative or more positive in a context-based model.

5 Conclusion

In this paper, we have constructed three models to assist in answering the question ”Would lyrics from Top 100 Billboard from 1965 to 2015 still reflect a similar distribution if we used a context-based neural network model and transformer model rather than a bag-of-words model”? After classifying the Billboard lyrics with our fine-tuned models, making the comparison to the LIWC probability distribution of the Billboard lyrics, and conducting the paired t-test we can we can conclude that the context-based model would predict a different distribution than the LIWC model for song lyrics. This could provide us enough evidence that when using a model that takes context into account, it results in a different probability distribution. This leads us to consider that the context-based model could potentially provide a more accurate representation of lyric sentiments.

The study follows the assumption that the model is accurate enough that the probability predicted are an actual reflection of the lyrics sentiments. It also followed the assumption that the user tagged valence was accurate as well. For next steps, to better improve on the model, a much larger sample size of song lyrics could be collected. The original MuSe data had a listing of 90K songs. Collecting and fine-tuning a BERT model on the complete size of the MuSe song listings and associated lyrics could assist in the accuracy of the model. Along with this, having a dataset’s valence scored by more professional artists rather than a general audience could assist in improving the accuracy of the model. As we described in the results section, there may be skewed bias or mislabeling of valence for many of the song lyrics. There could also be a challenge of interpretation for the users. To further improve the model, there is also potential to pre-train the model on the complete listings of the MuSe dataset rather than to only rely on the pre-trained Bert model. Overall, these improvements are more computationally expensive tasks and take more time not available in the timeframe for this study. Lastly, some music genres abbreviate their words or change the meanings of words entirely. This could be a challenge for most context-based models. To expand on the lyric processing for future studies, one could either eliminate lyrics without a consistent corpus match, or build a corpus that captures the meanings of these respective genres. In the future, we hope this study is improved upon as context-based models become more advanced. With the latest technologies, we could have a stronger understanding of the sentiment trend in popular music and what it depicts about our society.

References

1. Charlotte O. Brand, Alberto Acerbi, Alex Mesoudi (2019) “Cultural Evolution of Emotional Expression in 50 years of Song Lyrics”

2. Kaila C Putter, Amanda E Krause, Adrian C North (2022) “Popular music lyrics and the Covid-19 pandemic”

3. Christopher Akiki, M. Burghardt (2021) “MuSe: The Musical Sentiment Dataset”

4. Amy Beth Warriner, Victor Kuperman, Marc Brysbaert (2013) “Norms of valence, arousal, and dominance for 13,915 English lemmas”

5. Erion Cano, Maurizio Morisio (2017) “Music Mood Dataset Creation Based on Last FM Tags”