INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 8, Issue 4 - V8I4-1201)
Available online at: https://www.ijariit.com

Predicting stock market direction: hit songs' sentiment analysis by using bert

Sagar Bansal
sagarbansal099@gmail.com
Independent Researcher

Snigdha Patil
snigdhaspatil8@gmail.com
Independent Researcher

ABSTRACT

Various machine learning algorithms have been used to predict stock market prices and perform analyses. Investor sentiment has been taken into account to predict the prices. However, there are different ways in which investor mood or sentiment can be influenced. Studies have confirmed that people's mood or sentiment influences their choice of songs and likewise, songs influence people's mood or sentiment and in fact their buying behavior. This paper proposes a prediction of whether the stock market will be bearish or bullish on the basis of daily hit songs listened to by people on a popular music streaming platform. Only a few types of research have been conducted that propose a correlation between popular songs' sentiments and predicting the stock market. In this paper, the direction of the Dow Jones Industrial Average (DJIA) index is predicted from the lyrics of daily hit songs on Spotify listened to in the region of the United States of America. The model is trained on the dataset of lyrics of daily top 50 songs on Spotify from January 2017 to February 2022. The BERT (Bidirectional Encoder Representations from Transformers) model in Natural Language Processing (NLP) has been used to predict the direction of the DJIA index for the next day.

Keywords: *BERT Model, Stock Market Direction, Natural Language Processing, Sentiment Analysis*

1. INTRODUCTION

Stock market price prediction is difficult and not accurate, and hence, not reliable. However, predicting the direction of the stock market index can help people make better decisions about buying or selling stocks. Sentiments analyzed from the lyrics of the hit songs can be modeled as temporary but a genuine snapshot of the public mood. There have been studies that confirm the influence of people's mood on song choice and songs on the public mood. Researchers have confirmed that hit songs' sentiments harness people's moods and can be used to predict the stock market. Considering this, we decided to research this further using the BERT model for daily hit songs streamed on Spotify – a popular digital music service.

1.1 Why is the BERT model used for the sentiment analysis task?

TextBlob and Support Vector Machines (SVM) in Machine Learning are two popular techniques to perform sentiment analysis. However, sentiment analysis is an ever-evolving research field with many different real-life applications. The problem with these approaches is that it is difficult to correctly predict the sentiment of sarcastic words. The contextualized meaning of some words can mislead the predictions of the NLP model. Another is the Word2vec technique in NLP to generate word embeddings, that is, to learn associations between words from a huge text data corpus. This model can be used to detect synonymous words and for completing partial sentences. It represents each word as a vector of real numbers. The similarity between different numbers in the vector indicates the semantic similarity between the represented words. However, the issue with Word2vec is that it creates fixed word embeddings. For instance, consider the two sentences - "The Alchemist is a nice novel." and "The electric bulb was a novel invention.". The word 'novel' in both sentences holds a different meaning. Since Word2vec creates fixed word embeddings, it will fail in cases like the aforementioned. We want to capture the contextualized meaning of a word i.e., look at the whole sentence to generate number representation for a word and BERT does contextualized embeddings. For instance, the word embedding of "The electric bulb was a novel invention." and "The steam engine was a new invention." will be very similar. Likewise, "The Alchemist is a nice novel." and "Inferno is a nice book" will have similar embeddings. To overcome such difficult tasks, we thought of using the transformer-based BERT model.

2. RELATED WORK

Rachel Harsley et al. [1] explored the relationship between the sentiment of lyrics in Billboard Top 100 songs, the Dow Jones Industrial Average (DJIA), and a consumer confidence index. They hypothesized that the sentiment of Top 100 songs could be representative of public mood and correlate to stock market changes as well. They evaluated the correlation between the song polarity using the Pearson Correlation Coefficient and t-tests. The results indicated that the polarity of song lyrics and DJIA have a significant negative correlation. Although the absolute value of the correlation coefficient is half that of the baseline, that indicates the association is not as strong.

Adrian Fernandez-Perez et al. [2] developed a measure of investor sentiment based on the valence (the musical positiveness conveyed by a track) of Spotify songs that people listen to. They measured the average happiness of the songs played over seven days in a country and compared it with stock market trends over that week. Their research suggested that more-positive listening choices were significantly correlated with stock price gains. Furthermore, the same analysis was conducted for 39 other countries, and the results were the same.

While past work has looked at correlations between song polarity and public opinion using Pearson's correlation, and investor mood using song's valence, our work uses a Deep Learning approach to determine the direction of the stock market based on the overall sentimental analysis of daily hit songs using the BERT model.

3. DATASET

3.1 Storing and fetching data

Spotify's daily top 200 songs lyrics data from January 2017 to July 2021 are fetched from Kaggle. For the remaining days till February 2020, we scraped the song data from Spotify's daily hit songs US charts. And then lyrics for these songs are scraped from genius.com. The final dataset consists of daily top 50 songs because the song might repeat in the top 200 list every day at some rank and the dataset will have repeat values. For instance, if the rank of a song is 126 today, then the rank of the same song can be 98 tomorrow and both ranks fall within 200. So, to get more diversity of songs only the top 50 songs are considered in the final dataset. We created a label called target which has a value of 1 or -1, where 1 indicates the market is bullish and -1 indicates bearish based on the value of the DJIA index for the previous dates because today's songs will influence tomorrow's index. The value of the label is calculated from the DJIA index value obtained from Yahoo Finance.

3.2 Data pre-processing

We cleaned data by removing stop words and duplicate verses in the song's lyrics. Further, we removed the non-English words to achieve better accuracy. Next, we tokenize all of the sentences and map the tokens to their word IDs with the help of the BERT tokenizer. For each sentence, the below steps are followed -

1. Tokenize the sentence.
2. Prepend the `[CLS]` token to the start.
3. Append the `[SEP]` token to the end.
4. Map tokens to their IDs.
5. Pad or truncate the sentence to `max_length`
6. Create attention masks for [PAD] tokens.

[CLS] and [SEP] tokens are artificial tokens that are respectively inserted before the first sequence of tokens and between the first and second sentences. In order to analyze the data, we calculated the lengths of these sentences for all song lyrics and studied the percentiles (shown in Table-1). The plot of the density of sentences versus the token count. is shown in chart 1.

Table-1: Token counts

Measure	Token count
Maximum sentence length	1918
Average Length	576.26
50th Percentile	535.0
90th Percentile	899.80

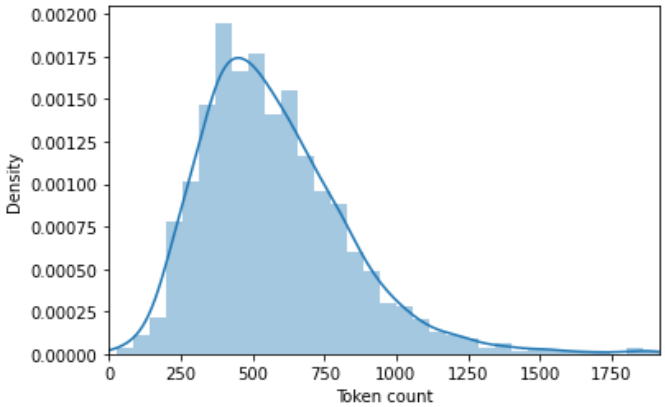

Chart-1 Density vs token count

4. MODEL ARCHITECTURE

BERT (Bidirectional Encoder Representations from Transformers) [7] is a Natural Language Processing Model developed by Google Researchers. There are two versions of the model $BERT_{BASE}$ and $BERT_{LARGE}$ that are trained on very large datasets. BERT model from transformers library – a Natural Language Processing library by hugging face is used for training. The song data for each date is split into test and train data. BERT Model tokenizer pre-processes the lyrics of each song and this pre-processed text is passed to the BERT model for fine-tuning. The Lyrics Sentiment Classifier uses the BERT model and a special mapping layer at the end of the model that maps BERT output 768 ($BERT_{BASE}$ model) to 1213 (number of unique days since we predict the index for each day). Then the linear classifier layer classifies this text into 1 or -1, where 1 indicates the market will be bullish and -1 indicates the market will be bearish.

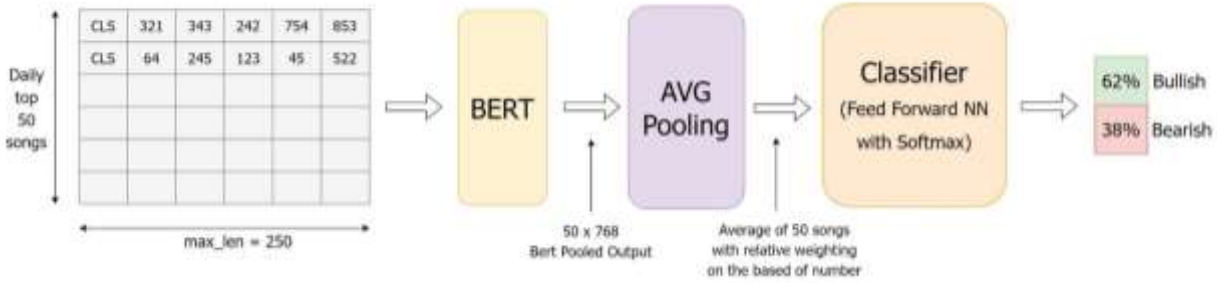

Image -1: Lyrics classification Model

5. RESULTS

We used AdamW optimizer and CrossEntropyLoss as our loss function. Then we fine-tuned the model for 5 epochs with a learning rate of 2e-5 to combat over-fitting. Data was fed into the model in a batch size of 32. Further, we used powerful Amazon EC2 DL1 instances powered by Gaudi accelerators from Habana Labs to train the model. This provided up to 40% better price-performance for training deep learning models compared to current generation GPU-based EC2 instances. After transfer learning, we saw a gradual decrease in the loss function (shown in chart 2) and an increase in the training accuracy of 90% and validation accuracy of 56% until epoch 3.

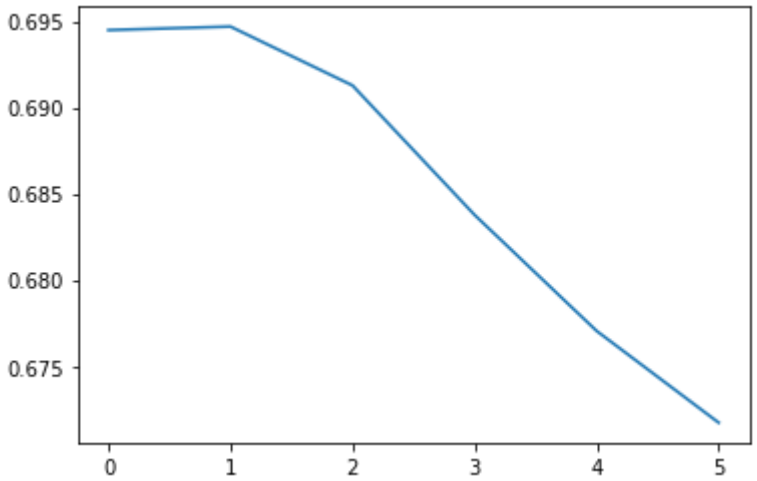

Chart-2 Loss function

6. CONCLUSION

In this paper, we explored whether stock market direction can be predicted by doing sentiment analysis on the daily hit songs listened to by people on Spotify. The Deep Learning approach of fine-tuning the BERT classifier has shown some positive outlook toward the hypothesis that the stock market can be driven by people's choice of music. This also bolsters the past research that explored the relationship between the sentiment of songs and the Dow Jones Industrial Average (DJIA), and a consumer confidence index. However, we believe there is a scope for future work to improvise the accuracy of the model. Some songs in the list of the top songs were not in English, but in other languages like Spanish and French. So non-English songs need to be translated into English to achieve better accuracy. Moreover, we aim to train the model to accommodate negations and modifiers around the sentiment word to determine the correct sentiment of the song.

7. REFERENCES

[1] Rachel Harsley, Bhavesh Gupta, Barbara Di Eugenio, and Huayi Li. Hit Songs' Sentiments Harness Public Mood & Predict Stock Market. In Conference: WASSA 16.

[2] Alex Edmans, Adrian, Fernandez-Perez, Alexander, Garel, Ivan Indriawan. Music sentiment and stock returns around the world. Journal of Financial Economics Volume 145, Issue 2, Part A, August 2022, Pages 234-254.

[3] YUNQING XIA, LINLIN WANG, and KAM-FAI WONG. Sentiment Vector Space Model for Lyric-Based Song Sentiment Classification. International Journal of Computer Processing of LanguagesVol. 21, No. 04, pp. 309-330 (2008)

[4] P. J. H. Daas and M. J. Puts. 2014. Social Media Sentiment and Consumer Confidence. Statistics Paper Series, No. 5. European Central Bank.
[5] Gordon C. Bruner. 1990. Music, Mood, and Marketing. Journal of Marketing, 54(4):94
[6] Andrea Chiorrini, Claudia Diamantini, Alex Mircoli, Domenico Potena. Emotion and sentiment analysis of tweets using BERT. Published in EDBT/ICDT Workshops 2021.
[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805
[8] Jiang Zhong, Yifeng Cheng, Siyuan Yang, and Luosheng Wen. 2012. Music sentiment classification integrating audio with lyrics. Journal of Information and Computational Science, 9:35–44

