Lyrics Analyzer

Amit Sharma, Jaspreet Singh, Deepa Vyasabhat, Ambika Na
Charles W. Davidson College of Engineering @ San Jose State University,
One Washington Square, San Jose, CA 95192-0080

Abstract - **Lyrics Analyzer is a tool developed for people associated with the music industry, especially the lyricists. It works on the model of training upon the historical data of already popular songs lyrics and through natural language processing extract their key features to prepare a dataset, on which various machine learning algorithm is trained and then applied to generate a prediction model for guessing popularity of new songs based on keywords extracted from the lyrical dataset.**

Index Terms - Lyrics Analysis, Natural Language Processing, Machine Learning, Dockerise, Python, Lyricist, Cloud Deployment, AWS

I. INTRODUCTION

This paper is about our final report on our Project - '***Lyrics Analyzer'*** which was done under the guidance of Prof. Rakesh Ranjan at San Jose State University, California. The tool is developed using the NLP and ML which are defined as :

"Natural language processing (NLP) is one area of artificial intelligence using computational linguistics that provides parsing and semantic interpretation of the text, which allows systems to learn, analyze, and understand human language." [1]

"Machine Learning (ML) is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves." [2]

II. RELATED WORKS

Capitalizing on the initial idea, we searched if there was any previous work already done with similar goals. We could not find anything with the same focus, although there was a paper [3] which was about a somewhat similar idea. We had a fair idea that in the real world, the popularity of a song is dependent on many factors i.e popularity of the performing artist, releasing label, composition, release platform, music genre, and so on. This means that any prediction based solely on song lyrics will not yield a very high degree of accuracy.

This is further supported by the paper described earlier which states that usually any non-neural lyrical classifiers will only able to achieve classification accuracy to near about just over 50% using machine learning techniques like SVMs, k-NN, and NB which are the most widely used techniques employed during the previous lyrical classification researches.

III. CORE IMPLEMENTATION

The two datasets were used while pursuing this project and are available at the Kaggle website [4] [5]. The first dataset contains the lyrics for all the popular songs that were listed in Top 100 songs for each year from 1964 to 2015 by songs ranking website Billboards Charts. So this dataset provided us with very accurate data and we labeled all the songs in this dataset with a target of one.

For the second dataset, as there was no dataset available which contained purely flop songs and due to lack of any other alternative we had to randomly select songs from this dataset, still we tried to include mostly flops and few average/ semi-hits songs and labeled them as zero to differentiate them from popular hit songs. Due to this, our accuracy was impacted but still, we were able to improve the overall accuracy of our project as compared to any of the previous research based on this subject and data.

But, the data contained in these datasets needed further processing, because this data was natural language data and we had to transform it using NLP techniques to extract features from these merged datasets. So, after completing this we had intermediate datasets that could be used as input for machine learning algorithms for training and testing a prediction model.

NLP and Feature Selection:

Total Length: The overall length of the song.
Unique Words: The number of unique words used in the songs
Repetition Factor :

1. Most Repeated Word in a song
2. Average Repetition of top 5 repeated words in a song

Usage of Top Words: We compiled the list of top 1000 most repeated words in the dataset and formulated a weighted index on it.

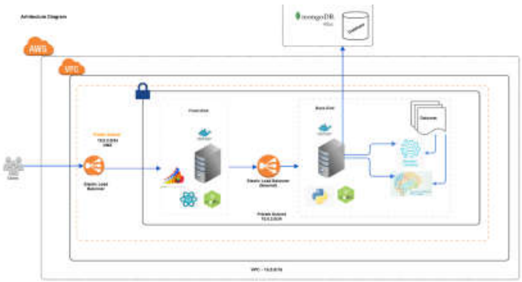
1. Weighted Index for the overall length of the song
2. Weighted Index for the number of unique words in the song

There were 7 columns overall with 6 features and a target variable 0 or 1. We had normalized the dataset also by calculating the mean of the dataset to account for the variability of the dataset. All the features were normalized against the calculated means and created another normalized dataset.

Machine Learning:

There were 2 datasets now (one with absolute features and other with normalized features). We decided to use both the classification and regression techniques on these datasets. We applied k-Nearest Neighbours Classification, Random Forest Classification, Random Forest Regression, Naive Bayes, Decision Tree Classification and Logistic Regression to train a prediction model. We were able to achieve an accuracy of about 65 to 70 % on all these trained models, which is quite reasonable as per the availability of datasets and the nature of data. Our core application/model was ready to predict the unknown lyrics written by the user and predict its popularity based on the same key features like Length of the song, Repetition Factor of words, and Usage of top words.

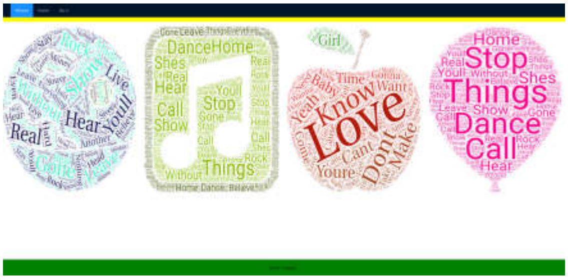
IV. ARCHITECTURE



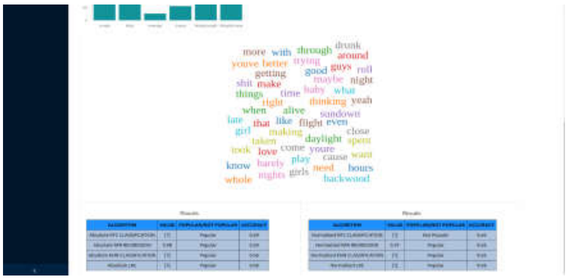
V. FRONT END

The frontend is based on react.js. We are including the AntDesign library for the CSS part. The welcome page includes word cloud visualization of most used words in popular songs. Users can register themselves and also can check their past searched lyrics. We are also providing visualization to users based on the cleaned dataset which will help them in deciding which words to include while writing lyrics. It will also help them in gaining insights into the structure of popular songs.

The welcome page has options for user login and registration.

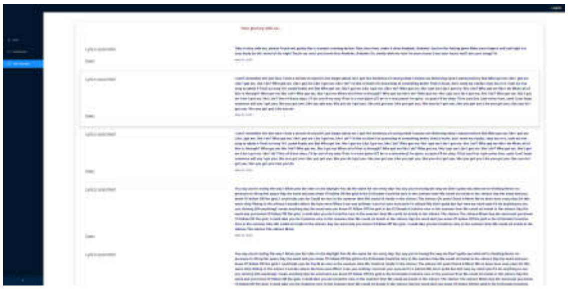


User main page where the user adds lyrics of the songs for which he requires predictions. NLP and ML algorithms run on this input and populate the prediction result with extracted features on this page.



The dashboard shows extracted features from the transformed dataset. This helps users to get insight on words and features to use for hit songs.

User journey page shows the history of lyrics predicted for particular user



VI. DEPLOYMENT

This web-based application is hosted on the AWS cloud. We have used docker version 2.2 to dockerize our frontend and backend node applications.

These docker images are also available in the Docker hub registry. They have been pulled and deployed on two different EC2 instances on AWS in a private subnet. We have used an external cloud-based MongoDB in the Mongo Atlas Cloud to store user-related information like authentication details and user history. Our backend points to this DB server. The back end is provisioned behind an Elastic Load Balancer so that we can scale the instances when needed based on user load. The same is done with the front end. Users can use the front end load balancer URL to access the application.

VII CONCLUSION

To sum up, there is great potential for more focused research on this topic, we were able to achieve better accuracy than the previous research but it is still far away from the expected accuracy of above 95 % to be a successful and consistent model. The key reason behind this low accuracy is because the popularity of a song depends upon many other variables apart from the songs' lyrics. The factors like music genre, the popularity of performing artists, the target audience, and so on are also equally important in determining the fate of a song. The other reason is the lack of availability of large amounts of labeled data for popular and flop songs with lyrics. With more research on this topic, and combining better datasets and enhancing the scope of the number of features used and sentiment analysis will yield much better results with higher accuracy.

VIII REFERENCES

1. https://www.ibm.com/watson/natural-language-processing
2. https://expertsystem.com/machine-learning-definition
3. https://www.irjet.net/archives/V5/i4/IRJET-V5I4578.pdf
4. https://www.kaggle.com/rakannimer/billboard-lyric
5. https://www.kaggle.com/mousehead/songlyrics