Comparative Analysis Of Music Streaming Platforms

Authors(**Team 9**)

Akhilesh Tawde
Raj Phadke
Rohit Jain
Rishika

Introduction

Spotify, Amazon Music, Pandora, & Apple Music being the world's largest music and video streaming services, allow users to stream music to their devices on demand.
They also provide the user, liability to provide positive or negative feedback for songs chosen by the service and based on the feedback, a subsequent selection of songs is recommended back to the user. All of these applications have variants, basic and premium.
On the other hand, Twitter is a news and social networking service on which users around the world interact using tweets. Users can post, tweet, retweet and like tweets. In recent times, Twitter is also important for business looking to expand their social reach. Using tweets, businesses can reach potential customers outside of the scope of normal marketing methodologies. Twitter hashtags serve the main purpose of these marketing schemes by helping businesses find users and conversations around it.
The overview of this project is to perform a sentimental analyst to check whether the business values of this music streaming platform are in line with voice of customer

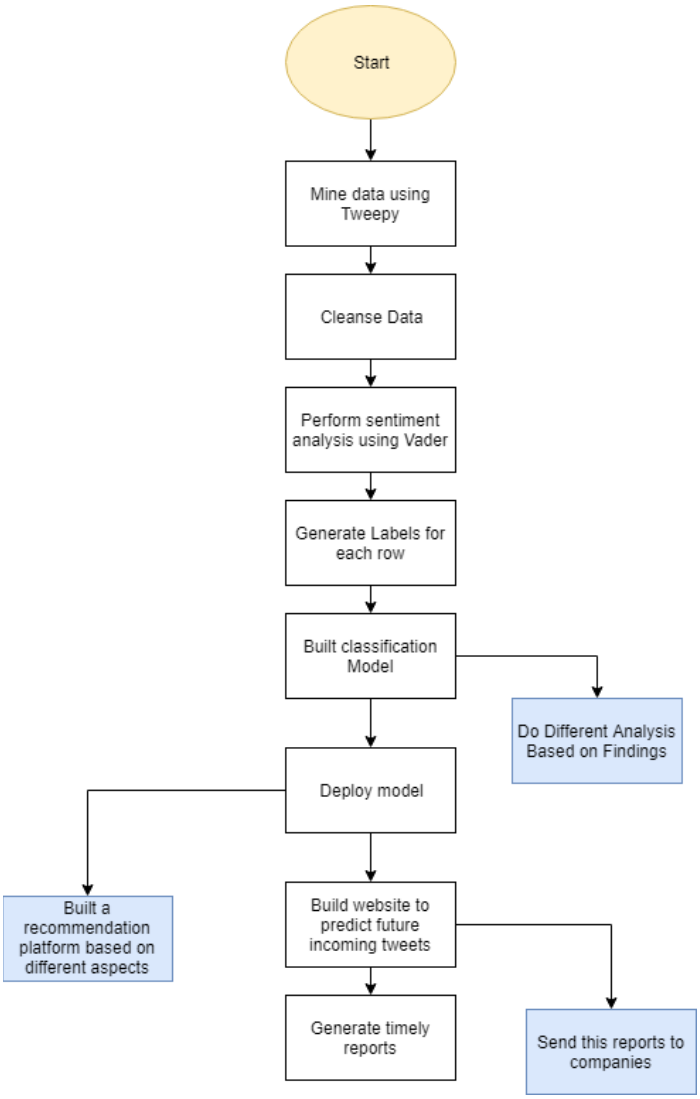
Project Goals

1. Analyse the tweet data and Google Playstore reviews to determine if the sentiment is positive/Negative/neutral and Perform Emoticon based Analysis
2. Compare different music platform based on the outcomes of sentimental analysis and more detailed aspects
3. Performing comparative analysis between different music streaming platform based on Cost, Performance, Advertisement and Stability.
4. 4. Build a website that hosts a dashboard showing analysis of these music streaming platform and it will have an input text box where the user can check whether the comment he/she is planning to post is a negative comment or a positive comment.Additionally generate a report by mining freshly prepared data from twitter in csv format and send the sentiment details to the provided user.
5. Generate a UI to choose between different parameters such as cost, advertisements, song collection, stability etc which will recommend the user best streaming platform according to its needs.

Technologies used

1. Mining data using twitter - Tweepy
2. Scraping - Beautiful Soup, Selenium web driver
3. Sentiment Analysis - Vader
4. Classification Model - SVM (Support Vector Machine) , Naive Bayes
5. Pipelining - Flask
6. Analysis - Tableau
7. Cloud Hosting Platform : Heroku

Pipeline Design



Implementation Details

Web Scraping and Data Mining

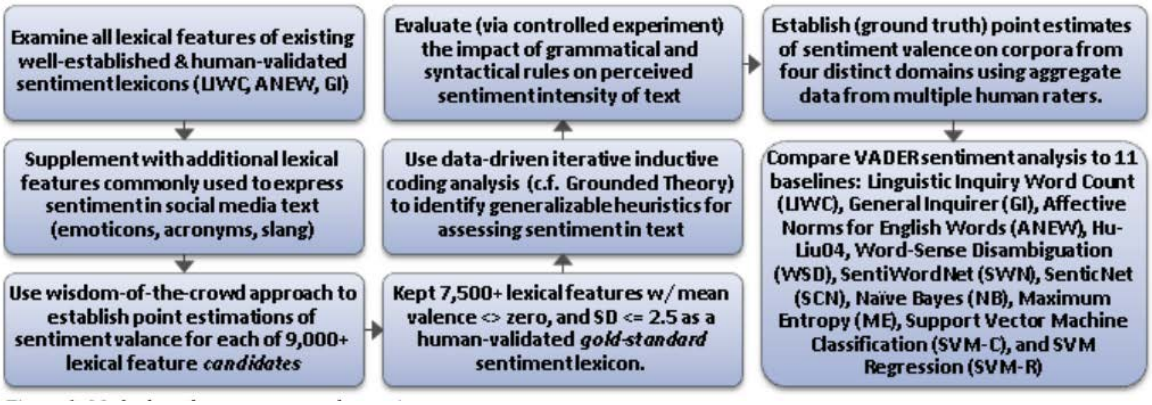
Web Scraping - Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on.
We used Beautiful Soup and Selenium Web Driver to scrape data from Google Playstore.

Data Mining - Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use.
We used tweepy api to mine data from Twitter.
Tweepy : The API class provides access to the entire twitter RESTful API methods. Each method can accept various parameters and return responses. For more information about these methods please refer to API Reference.

Sentiment Analysis Using Vader

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative.

VADER has been found to be quite successful when dealing with social media texts, NY Times editorials, movie reviews, and product reviews. This is because VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.



Advantages of using VADER

VADER has a lot of advantages over traditional methods of Sentiment Analysis, including:

- It works exceedingly well on social media type text, yet readily generalizes to multiple domains
- It doesn't require any training data but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon
- It is fast enough to be used online with streaming data, and
- It does not severely suffer from a speed-performance tradeoff.

Model Development Using Tagged Dataset

Obtained dataset from vader sentiment analysis and used it as a training dataset on which we applied various classification models to predict the input given to the model as positive, negative or neutral.

Model Development for Recommendation Platform

We used bag of approach to create 8 bags which included factors like good recommendation, cost, advertisements and stability. So based on previous users and their respective comments, we tried to run the reviews against our bags. The category against which we got maximum hits for a review, we classified that review in the same category. These observations were used as a training dataset. For recommending a music platform to the user when he made demands like music platform being inexpensive yet stable etc, we used this training dataset.

Emoji Analysis

Emoticon Analysis : The goal of **sentiment analysis** is to define automatic tools able to extract subjective information, such as opinions and sentiments from natural language texts. When dealing with contents coming from social media, like Facebook and Twitter, we need to leverage new types of contents and diffusion models that need to be modeled explicitly starting from the language.
By careful observation of data we concluded that people often use emojis to express their sentiments rather than using exact words. So we believed there was a need to do an emoticon analysis. So we extracted the emoticons from the scraped data.These emoticons were in form of symbols. These symbols were specific to every emoticon. Also we created a csv which portrayed the sentiments related to these emoticon's symbols very clearly. We ran our collection of emoticons against this csv and analyzed the results into two categories - happy or sad.

Website with multiple webpages

Section 1

First web page consists of a text box which takes input from the user and predicts the sentiment of the entered comment or review by categorizing it into positive, negative or neutral sentiment.

Section 2

Secondly web page consists of 8 check buttons which will allow user to input his/her preferences about the musical platform. Then according to the input, user will receive a recommendation about his needs/ requirements.

Section 3

This web page follows a pipelined process where user can enter three inputs :
1. Gmail credentials which include username and password -
2. Hashtag # inputs - The tweets containing the hashtags will be mined.
3. Email Id's - The generated report will be sent to the respective email id entered by the user if requested.

Section 4

Presenting different kinds of analysis based on the data which were created on Tableau.

Analysis of the Model

1. Prediction of sentiment of incoming tweet/comment

By using vader sentiment analysis we generated the output which tagged the reviews as positive, negative and neutral. By using this tagged data as our training set. We divided our data into features and labels where the features were the reviews collected from the users and label as a tagged output
Using this training set as an input we created various classification models where the input were the reviews/comments/tweets from the users and the predicted output was positive, negative or neutral.

We created 7 classification model which are SVM, Logistic Regression, Bagging, Boosting , Random Forest and Naive Bayes. Then we compared them on the basis of their MAPE, RMSE, MSE and accuracy. We obtained best accuracy ie 73 % by implementing SVM so we decided to proceed further with the same model. Then we hypertuned the parameters of SVM which increased the accuracy by 1 or 2 percent.

2. Development Of model for recommendation platform

We used bag of approach to create 8 bags which included factors like good recommendation, bad recommendation, good cost, bad cost, stability and instability. So based on previous users and their respective comments, we tried to run the reviews against our bags. The categories against which we got maximum hits for a review was classified in that category. All this was used a training dataset. Then used for recommending a music platform to the user when he made demands like music platform being inexpensive yet stable etc.
We developed 5 models such as Naive Bayes, Random Forest , SVM , Logistic Regression.Naive Bayes gave us the best results by hypertuning the parameters. We also tried to implement auto ml h2o which gave us similar accuracy more or less.

Details on how to run the model

1. Prediction of sentiment of incoming tweet/comment

- After obtaining the vader sentiment output, we will divide the dataset into two parts as features and label. Use reviews as features by putting them in features and the labels are sentiments obtained from Vader Sentiment Analysis.
- Import the package named SVM from scikit learn and use features as “x” and labels as “y” for the model.
- Fit the model by considering these values of “x” and “y”
- Predict the output of incoming reviews/ comments/ tweets ie sentiment in our case.

2. Development Of model for recommendation platform

- After implementing bag of words and running this against our dataset, we will obtain the frequency of words.By using these 8 columns as our features, our label would be the platform(Pandora, Spotify, Amazon Music) the user is currently commenting on.
- Import the package named Naive Bayes from scikit learn and use features as “x” and labels as “y” for the model.
- Fit the model by considering these values of “x” and “y”
- Recommend a suggestion on which music platform to use based on the input paramters user selects.