Initial Project Plan

Author - Darren O’Callaghan

Supervisor – Daniela Tsaneva

CM3203

One semester individual project – 40 credits

Project Title

Predicting a Musical Artists Spotify success based on the Sentiment of their Social Media presence.

Project Description

Trends in the music industry have a been one of its most important aspects since its commercialisation. In more recent times, it has been noted that on-demand music services have become the most popular way to access and listen to music [1]. These services produce and publish huge quantities of data that can be analysed to investigate trends in the music industry in modern times. Additionally, social media services have also grown by huge amount in recent times, as has their user base, and the therefore, so has the data that is available from them. Many people on social media tell their friends and followers what they are currently listening to, and this has even been used by some to create new datasets based on listeners who have common listening habits. [2] Social media users also describe their feelings on many topics, and these tweets have been explored in various ways with Sentiment analysis [11]. In my project, I plan to explore the relationship between the general sentiment of musical genres and artists that exist on social media, and their streaming performance on Spotify.

To explore this relationship, I will need to collect relevant data from twitter (to then determine the general sentiment around different musical subjects) and from Spotify, to determine the subjects streaming performance. For twitter, I will first try to use their streaming API [3]. Twitter must approve use of this API, so I have applied to use it for my project. If the number of tweets obtained is not enough for this project, I plan to investigate using other open-source tools to collect data like *Twint* [4] or *Tweepy* [5]. To collect Spotify data, I plan to create my own *Selenium* [6] scraper and collect the information from spotifycharts.com, which is maintained by Spotify, and has their chart history going back several years. I will likely store both the Spotify and twitter data in either a MySQL or MongoDB database (I will complete further investigation into which will be most effective to use for this project), to then use when conducting my sentiment analysis and when analysing the relationship between the sentiment and Spotify chart performance.

For conducting the sentiment analysis, I will likely use the *Spacy* [7] python package, to process the tweets that I have collected, and I plan to investigate using Spacy’s built in classifier or the scikit-learn package [8] for building the actual classifier. This is because both seem to relatively accessible for someone who has limited experience with machine learning. If both provide limited results, or I have time during my project, I am open to investigating using other tools for the sentiment Analysis/machine learning phase of my project. I will then conduct additional machine learning with the aim to produce some form of a regression, that can predict streaming performance from a given determined social media sentiment.

Project Aims and Objectives

I wish to study how different types of Sentiment that fans have for an artist effect their commercial performance (represented by total Spotify streams in this project). Key questions I have are:

- Does positive sentiment lead to longer term success for a given artist, particularly when they publish a new single or Album?
- Is negative sentiment truly bad (surely a listener would have to stream a song to know they do not like it). Do songs, artists, or whole genres with negative social media sentiment struggle to maintain strong streaming numbers over time? Or is the term “all publicity is good publicity” true?
- What is the effect of neutral sentiment? Does it suggest that nobody is listening to a particular genre or artist- or are more?
- Do the same relationships between sentiment and streaming performance hold for extremely popular artists, and those with a smaller following?
- Do tweets that describe sentiment towards an artist **but not their music in particular** have the same relationships to streaming performance.

To effectively answer these questions, I plan to:

- Collect streaming statistics that have been made public by Spotify, which are available at https://spotifycharts.com/
- Collect relevant tweets from discussion on Twitter around musical Artists and Genres.
- Create a Sentiment Analysis classifier to accurately determine the general sentiment that Twitter users feel towards these artists and genres.
- Use my collected sentiment, and Spotify data to discover patterns between the general sentiment around an artist or genre, and its streaming performance- likely through additional machine learning.

Ethics

An important point when thinking about the ethics of this project is around ensuring the data that I collecting is free to use for academic purposes. My two major sources will be Spotify and Twitter. The Spotify data I intend to use resides at spotifycharts.com, and its associated robots.txt file indicates that the data available there is free to be collected by any means [link robots.txt]. For Twitter, their streaming API has been used in several other areas of research that I have investigated at the start of my project. There are also several open-source 3rd party tools for accessing Twitter data which I may use; if these are used, I will endeavour to ensure that my project abides by Twitter’s Terms of Service throughout. To the best of my knowledge, both sources investigated at this stage (Spotify Charts page and Twitter API) are free to use and collect data from.

Work Plan

Please see the associated Gantt chart that I have submitted alongside this report. I plan to update this Gantt chart throughout my project, as well as share and review it with my supervisor on a regular basis.

Major Milestones

The major milestone deadlines for this project are as follows:

1. Complete initial plan: 8/2/2021.
2. Complete Scrapper for Spotify data: 19/2/2021.
3. Collect required Spotify database into well designed database: 26/2/2021.
4. Collect Twitter data on musical artists and genres: 10/3/2021.
5. Complete sentiment analysis on Twitter data to obtain general sentiment on musical artists and genres: 31/3/2021.
6. Complete machine learning analysis into Sentiment and streaming data relationship: 12/4/2021.
7. Complete final report: 10/5/2021 (leaving room for final checks and submission time)

Meetings with Supervisor

I also plan to meet regularly with my supervisor, Daniela Tsaneva, on a weekly basis generally. In these meetings we will review the work completed up until that point, documentation that I have recorded, and review a regularly updated version of the Gantt chart that I have started, and attached with this report.

References

[1] Liikkanen, L.A. and Åman, P., 2016. Shuffling services: Current trends in interacting with digital music. *Interacting with Computers*, *28*(3), pp.352-371.

[2] Pichl, M., Zangerle, E. and Specht, G., 2014. Combining Spotify and Twitter Data for Generating a Recent and Public Dataset for Music Recommendation. In *Grundlagen von Datenbanken*, pp. 35-40.

[3] Twitter, inc. 2021. *Twitter API Documentation.* Available at: https://developer.twitter.com/en/docs/twitter-api [Accessed: 6 February 2021].

[4] Francesco Poldi et al. 2020. *Twint Source Code.* Version 2.1.21. [Source Code] Available at: https://github.com/twintproject/twint/tree/master [Accessed: 6 February 2021].

[5] Pablo Rivera et al. 2020. *Tweepy Source Code.* Version 3.10.0. [Source Code] Available at: https://github.com/tweepy/tweepy [Accessed: 6 February 2021].

[6] Software Freedom Conservancy et al. 2018. *Selenium Source Code.* Version 3.141.59. [Source Code] Available at: https://github.com/SeleniumHQ/selenium/ [Accessed: 6 February 2021]

[7] ExplosionAI GmbH et al. 2021. *spaCy Source Code.* Version 3.0.1. [Source Code] Available at: https://github.com/explosion/spaCy [Accessed: 6 February 2021]

[8] The scikit-learn developers. 2021. *Scikit-learn Source Code.* Version 0.24.1. [Source Code] Available at: https://github.com/scikit-learn/scikit-learn [Accessed: 6 February 2021]

[9] The scikit-learn developers. 2021. *Scikit-learn Source Code.* Version 0.24.1. [Source Code] Available at: https://github.com/scikit-learn/scikit-learn [Accessed: 6 February 2021]

[10] Spotify AB. 2021. *Spotify Charts.* Available at: https://spotifycharts.com/robots.txt [Accessed: 6 February 2021]

[11] Nausheen, F. and Begum, S.H., 2018, January. Sentiment analysis to predict election results using Python. In *2018 2nd international conference on inventive systems and control (ICISC)*, pp. 1259-1262. IEEE.