Università degli studi di Milano-Bicocca

Social Media Analytics Report

How TikTokers are shaping the music scenario in the USA

Authors:

Giacomo De Gobbi - 860913 - M.Sc. Data Science
Davide Vercesi - 852483 - M.Sc. Data Science
Daniele Monterisi - 853257 - M.Sc. Data Science

Contents

Abstract

This project analyzed the behavior of the top influencers on TikTok, focusing on the use of musical contents. Data was extracted with the aim of measuring the importance of the songs inside the platform. This investigation led to studying how homogeneous is the use of the musical products, analyzing the network between the top 200 most influential American TikTokers. The features of the most important and viral songs of the social network were studied both in terms of lyrics and in terms of sound characteristics. Lastly, the research led to an analysis into whether the songs that are trending on TikTok have a 'spillover effect' into other prominent social media platforms, thus becoming mainstream in the real world.

1 TikTok

In 2014, a social media app named Musical.ly (pronounced Musical-ly) became incredibly popular with the 13-18 years old demographic. The main purpose of Musical.ly was user-generated videos that combined popular songs with videos from the users (often called Musers). The most popular use of the app was to create videos where they were lip syncing and dancing. By mid-2017, the Musical.ly app had over 200 million users.
In 2016, Chinese app developer ByteDance created an app named Douyin, a rival to Musical.ly. Launched initially only in China, the app was renamed and rebranded to TikTok for better international appeal. Within a year, the TikTok app had more than 100 million users, and the popularity of lip sync videos continued to rise.
In late 2017, Musical.ly was acquired by ByteDance for a fee of $800 Million. In 2018, Bytedance consolidated the user accounts of Musical.ly and TikTok, merging the two apps into one under the name TikTok. With this unified brand and user base, the app began to increase in popularity very quickly. TikTok became the most downloaded app on the Apple App store in early 2018, surpassing Instagram, WhatsApp, and YouTube.
TikTok has seen an immense surge in downloads in 2020. It has exponentially grown since 2019, becoming the most downloaded app on the digital market, bolstered by COVID-19 lockdowns. Its popularity, like for the other huge social media networks, has overflown over the virtual barrier and it has also become a geopolitical player. For instance, TikTok has been banned by India along with 58 other Chinese-owned apps in July in response to escalating border tensions between the two countries. The Trump administration issued an executive order prohibiting TikTok and the Chinese messaging platform WeChat from conducting transactions in the United States, before voiding it, making way in September for Oracle and Walmart companies to take control of TikTok's

US operations.
But what makes TikTok so different? TikTok subverts the standard template of follower-based profiles. Operating on a communal homepage, users vertically swipe through an infinite stream of 15-second videos available on the app. It's an ecosystem where users can accumulate thousands of views on a video despite having zero followers, and it offers not only a more accessible point of entry than other apps. Unlike other platforms like Spotify, and Apple Music, which are strictly music sharing and listening platforms, contents on tiktok range from memes to viral punchlines and dances, using music as supporting soundtracks.

2 Data Acquisition

In order to obtain data from the TikTok platform a REST API was used. This consists of an application programming interface that implements a set of architectural principles referred to as Representational State Transfer (REST). More specifically, a Stateless interface was used since each request from the client to the server contains all the information necessary to understand the request and the session state is not stored on the server but exclusively and entirely on the client. In addition, the request from the client to the server also contains the authentication information stored in an HTTP Authorization header which is not stored by the server.
With the use of the API, the data and metadata relating to the content of the posts published by the 200 most followed TikTok users were requested. Moreover, each request contains a limit of posts set to 300. The requested information was obtained in JSON (Javascript Object Notation) format and was stored in a dataframe in order to proceed with subsequent manipulations. A variable number of posts was obtained for each user, and the following information was obtained for each post:

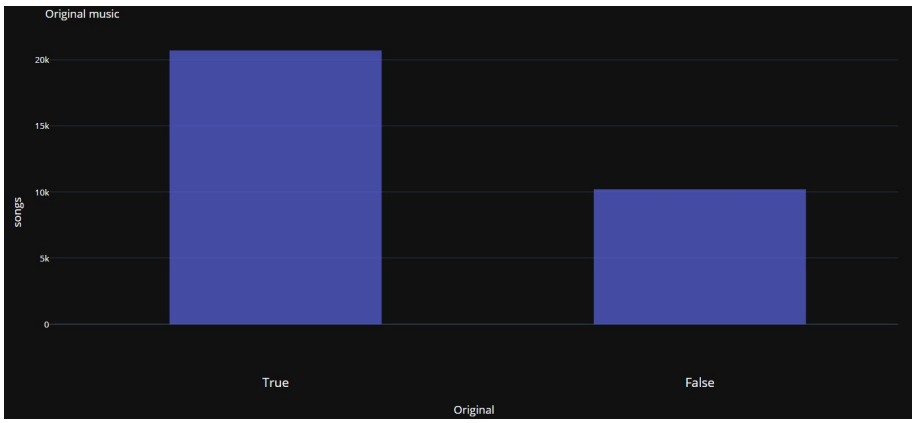
- `author_name`: username of the post creator;
- `create_time`: creation post timestamp;
- `music_author`: author of the song;
- `music_title`: song title used in the post;
- `original_music`: dummy variable indicating if the song was created from users or musicians;
- `diggCount`: number of likes of a post;
- `shareCount`: number of shares of a post;

- `playCount`: number of plays of a post.

The resulting dataset contains 33500 posts.

3 Data Wrangling

From this dataset only relevant information was filtered based on our domain expertise, in order to obtain only crucial information that would be significant to the declared scope of this project.
Firstly, users considered as not "pure TikTokers" were removed. This term refers to users who have gained popularity creating and sharing content mainly use TikTok as well as other related social networks.
As a result, all companies, actors and singers' accounts were dropped (e.g. Will Smith, Selena Gomez, H&M).
All posts with the attribute `music_original = True` were discarded as these contents do not contain songs by musicians but only consist of sound effects, part of speeches, remixes created by TikTok's users (non-copyrighted material). From the histogram below we can see the clear preponderance of these types of contents over the post with copyrighted songs.



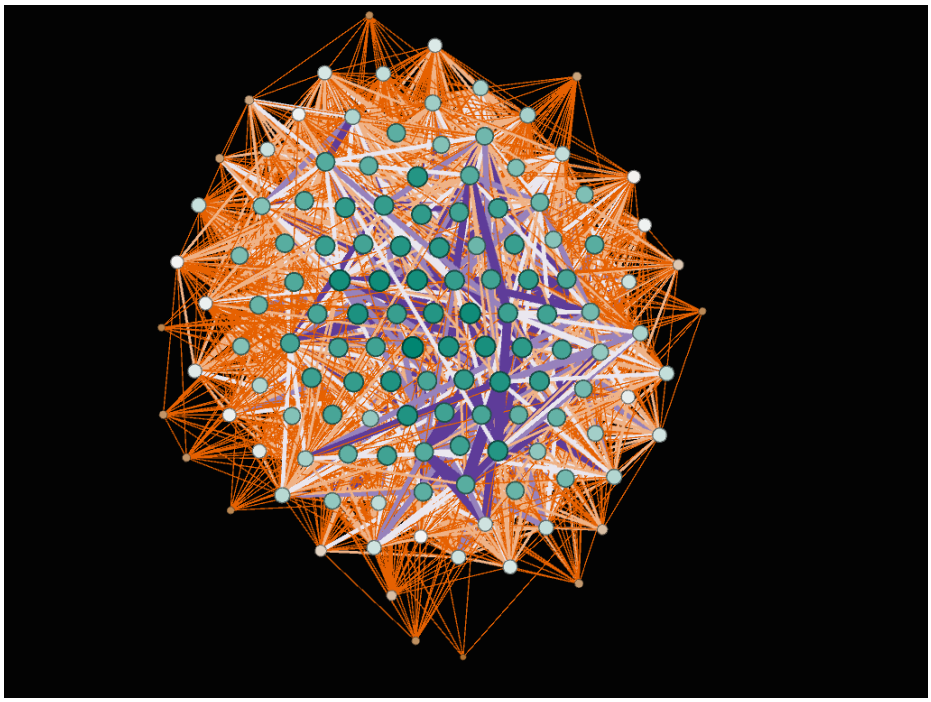
Data was then filtered in order to remain with posts that used songs that were released only in 2020 in order to produce an accurate trend analysis. Lastly, only posts that used songs that were utilized for at least ten different TikTokers were taken into account. This step was crucial in order to facilitate the sentiment analysis on the lyrics.

4 Network Analysis

For this analysis TikTok influencers were considered as nodes and the songs in common as edges. Additionally, the weight attribute was added as the number

of songs in common used by TikTokers.
The resulting graph has the following metrics:

- Number of nodes = 132
- Numbers of edges = 5687
- Average degree = 83.62
- Diameter = 3
- Density= 0.62



This visual representation created with the software Gephi is useful to understand the network. The colors of the nodes change from brown and white to green indicating an increase in the node's degree. On the other hand, the edges' color varies from orange/white to a darker purple indicating an increase of the edges' weight.
Additionally, the modularity measure was calculated using library **Networkx** from Python. The score was around zero, a result consistent with the metrics. As a result, the graph is significantly dense and without the presence of communities or hubs. This is undoubtedly interesting. It seems that the top influencers of the platform tend to use the same viral songs for their video contents.

5 Sentiment Analysis

Given the limited presence and use of textual contents within the platform of TikTok, a slightly different sentiment analysis was carried out from the 'standard' procedure normally applied on other social networks such as Twitter, Facebook, or Reddit.
So, instead of using comments or hashtag, the sentiment analysis was executed on the songs used by the top influencers in two different forms:

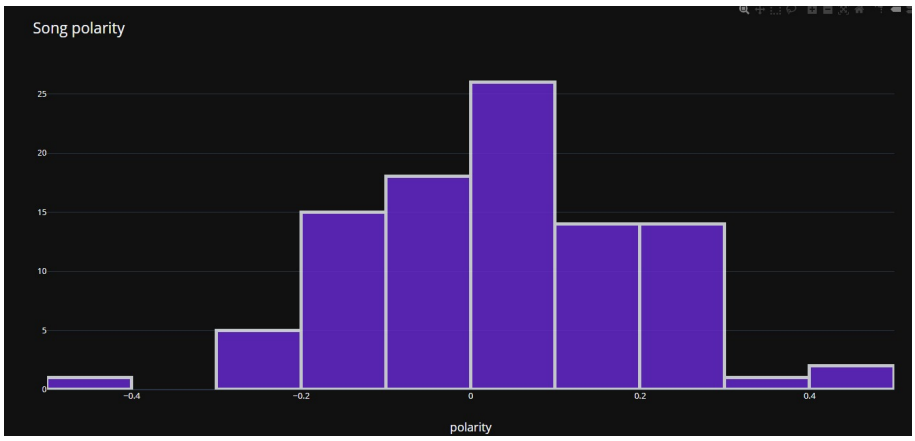
- Analysis of the lyrics of the songs (web scraping on Genius);
- Analysis of the characteristics of the sound (Spotify API REST).

5.1 Sentiment on Lyrics

Sentiment analysis on song lyrics was performed using the **TextBlob** library available for Python. Data was arranged in order to obtain better results and insights for the analysis.
The steps were the following:

- Turning text into lowercase, removing numbers and removing punctuation;
- Tokenization;
- Stop words removal.

The `PatternAnalyzer` command was used to obtain the polarity of each song. This ranged from -1 (negative) to +1 (positive).



As shown in the graph above, the polarity has a symmetric distribution with the mean around zero. The results are quite interesting. The expectation was an asymmetric distribution shifted towards positive values of polarity.

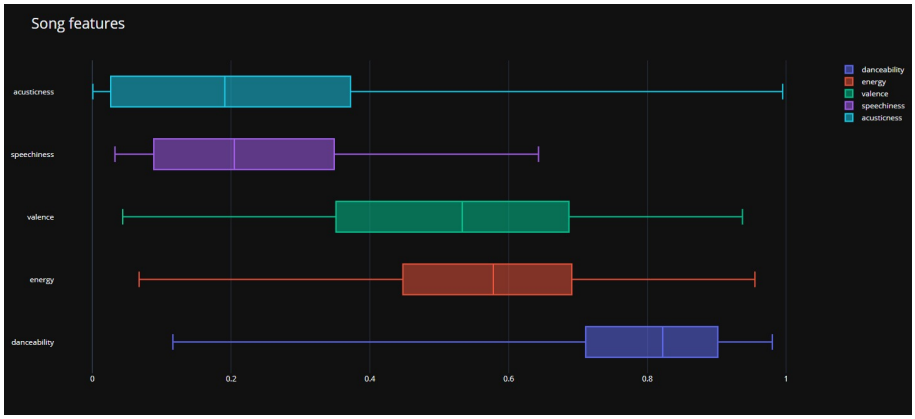
This prediction was supposed considering the reputation and perception of the social network by the public due to its light-hearted/easygoing content shared and the audience targeted.
A simple explanation to the obtained results could be by the time limit of 15 seconds for video content shared on TikTok while the sentiment analysis was carried out on the whole lyrics of these songs.

5.2 Sentiment on Sound Features

The sentiment on the sounds of the songs was performed using the **Spotipy** library available for Python. With the function `audio_features`, some metrics were extracted related to the sound characteristics of each song.
Specifically:

- **Acousticness**: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents with high confidence that the track is acoustic.
- **Danceability**: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **Energy**: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy;
- **Speechiness**: Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- **Valence**: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

As expected, TikTok and its users prefer songs with high energy and danceability that are more suitable for the creation of viral dances and lip-syncing

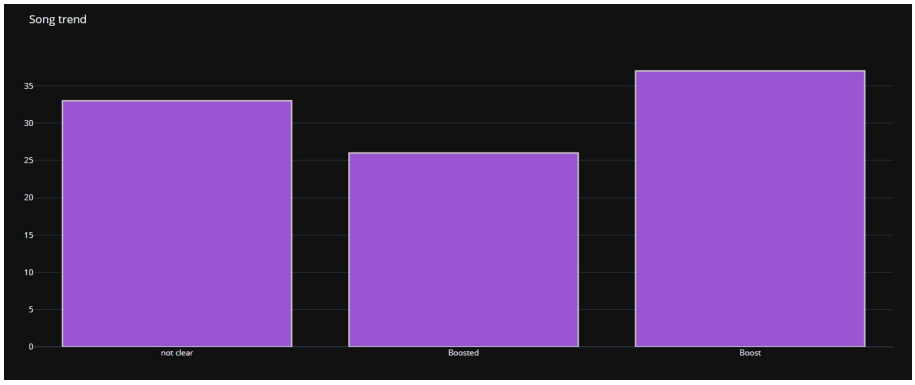


videos. On the other hand, songs with a preponderant acousticness and with speechness are rarely used, underlining the nature of the content predominantly shared on the platform.
The results for valence are quite surprising. A higher value was expected for this feature given the young target and the easy-going mood of the content on TikTok. The possible explanation of such a result could be related to the same reasons argued in the findings of the sentiment analysis of the lyrics above.

6 Impact Analysis

In order to assess if there was an impact of TikTokers on the music industry outside the platform the Google Trends data were exploited. More precisely, data related to the Google searches of the previously analyzed 96 songs was extracted with the help of the API, taking into account only searches made in the United States during 2020.
A function was created in order to classify songs into three categories:

- Boost: if at least 50% of the total *play count* for a song in TikTok occurred in a period before the peak week in Google Trends (score=100). It was considered that TikTokers have boosted the popularity of the song outside the platform;
- Boosted: if less than 50% of the *play count* for a song in TikTok occurred in a period before the peak week in Google Trends (score=100). It was considered that the song was trending before its use on TiktTok and the top users exploited its previous popularity to gain more views;
- Not Clear: there is a lack of data for the song in Google Trends or its evolution over time is difficult to interpret.



From the histogram above it can be easily assessed that for more than a third of the songs, TikTokers have boosted their popularity and virality outside the platform. This result is consistent with the vision of some journalists and musical critics of the unstoppable and extraordinary power of TikTok to dramatically shape and change the music industry.

7 Conclusions

Songs on TikTok are not the central element, but undoubtedly remain central to the content created, having a relevant impact on their popularity outside the platform. Given the results found, it can be easily predicted that in the near future the music industry may more likely induce artists, especially in the genre of pop and dance to produce songs that are "Tik-Tok friendly" in order to obtain major success from the public. For these reasons, major attention is recommended to the study of TikTok, monitoring its temporary trends.