Using social media to reveal social and collective perspectives on music

Deniz Iren, Cynthia C. S. Liem, Jie Yang, Alessandro Bozzon
Delft University of Technology
Mekelweg 4, 2628 CD Delft, the Netherlands
{y.d.iren, c.c.s.liem, j.yang-3, a.bozzon}@tudelft.nl

ABSTRACT

How can social media be used to reveal latent social and collective perspectives on music? Our work addresses this question by introducing a Twitter dataset surrounding the `Top 2000`, a yearly national broadcasting event in The Netherlands. The `Top 2000` is recognised as a valuable case study into the role of music as a social nostalgia-inducing phenomenon, triggering collective and autobiographical memories. Our dataset, containing enriched Twitter information over the `Top 2000` voting and broadcasting timeline in 2015, demonstrates how the broad audience support of the event enables data-oriented studies of the public response to and public significance of the aired songs.

CCS Concepts

•Information systems → Web searching and information discovery; •Human-centered computing → Social content sharing; Social media; •Social and professional topics → User characteristics;

Keywords

social media, data analysis, music preferences

1. INTRODUCTION

Where people are, music is consumed. In many cultures, music is an integral element accompanying significant social activities [4], and an essential part of persuasive multimedia such as commercials [2]. Also in private spheres, music accompanies us in daily life[1]. Our music taste has been influenced by our everyday environment and the social influence of our surroundings [1, 3]. 'Taste cultures', to which 'taste publics' subscribe, can be defined by certain sociodemographic characteristics [12]. In other words, our music taste can be an indication of who we are (or wish to be), and whom we identify with.

[1]see e.g. the recent Sonos consumer study at http://www.musicmakesithome.com/.

WebSci '16, May 22-25, 2016, Hannover, Germany

DOI: http://dx.doi.org/10.1145/2908131.2908178

In parallel, with respect to the Web, the strong social component of Web communication platforms like social media (e.g. Twitter, or Instagram) also might reflect—or reveal—complex socio-economic and socio-cultural aspects of the real world. Can social media data therefore be used to reveal latent social and collective perspectives on music?

Several datasets exist which consider social media as a resource for music consumption data. Two notable datasets based on information from the microblog service Twitter are the "Million Musical Tweet dataset" [10] and the "#nowplaying" dataset [15]. Both were created by crawling the Twitter stream using keywords (or hashtags) of a generic nature, and with strong connection to automated postings by music services (e.g. Spotify). Therefore, we cannot tell for sure in what context music was played, and whether there truly was an active human listener connected to the consumption.

For our current work, we seek social media data with higher certainty of social and collective significance of music consumption behaviour, and conscious choices by users to listen to the music and publicly speak about it. To this end, we consider microblogging activity surrounding the `Top 2000`, a yearly music ranking and broadcasting event of the 2000 "greatest songs of all time" in The Netherlands, which has grown to become an event of high national public, cultural and social significance and fascination [9].

The contributions of this paper are threefold:

- A novel dataset of microblog activity surrounding the `Top 2000`, enriched with further metadata about the voted and mentioned songs. We describe the employed methodology, and release the resulting (enriched) dataset;
- A study showing how the `Top 2000`-related social media activity of users reflects the voting preferences of the nation;
- An analysis of the mention behaviour of users during the airing of the `Top 2000`, showing how the dataset can be used to reveal properties of users relevant to the study of music consumption preferences.

The remainder of the paper is organised as follows: Section 2 briefly introduces the `Top 2000` event; Section 3 describes the data gathering and enrichment process used to create the dataset, which salient properties are listed in Section 4; Section 5 describes two studies performed on the novel dataset; finally, Section 6 concludes the paper.

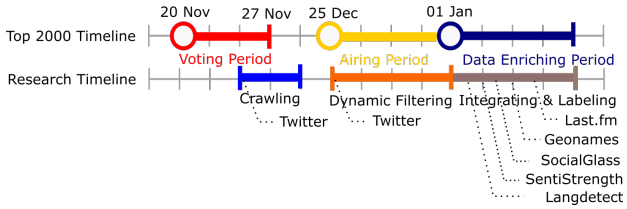


Figure 1: The `Top 2000` timeline, and the steps of our research.

2. THE TOP 2000

The `Top 2000`[2] is a ranking of the 2000 "greatest songs of all time" in The Netherlands, which since 1999 is yearly established through public voting. The 2000 most voted-on songs are aired on public radio in a marathon broadcast between Christmas and New Year's Eve. The broadcast is broadly followed by the Dutch population: in 2015, 8.1 million people (almost 50 % of the total Dutch population) were reported to have tuned in to it[3].

Given the broad social support for the event spanning multiple generations, the `Top 2000` has grown into a significant national broadcasting event that connects people [9] and reflects both collective and autobiographical memory [14]. Nostalgia plays an important role in the event, confirming existing findings in literature that autobiographical connections to songs are an important factor in inducing musical nostalgia [6]. Personal stories of people connected to songs in the ranking are explicitly featured in the broadcast.

In this paper, we consider the 2015 edition of the `Top 2000`, of which the timeline is shown in Figure 1. Between the 20th and the 27th of November, 2015, people could vote online by selecting up to 25 of their favourite songs. Voting lists could publicly be shared through social media. On the 16th of December, the final ranking was published online, along with an hourly broadcasting schedule for the entire airing period. Songs are played in a pre-defined order, counting down from the song at rank 2000 until the song at rank 1, which airs just before the turn of the year.

3. DATA COLLECTION & ENRICHMENT

For the data collection procedure, we focused on Twitter microblog posts during two phases of the `Top 2000`: the voting phase, in which people could nominate songs for the ranking, and the airing period, in which the actual ranking was broadcasted. As a consequence, the tweets represented in our dataset can be split into a *voting collection* and *airing collection*, for which a slightly different acquisition methodology was used. Besides, upon collecting the Tweet data, we performed several enrichment steps to gain richer information surrounding the Tweet content.

Voting Collection. As mentioned in the previous section, people voting on their favourite songs could publicly share their voting lists on social media. We consider posts with these lists to be of importance, as they reveal explicit public music preference information that can be related to users. We searched the Twitter timeline for the standardized URL prefix of voting pages (`"top2000.nl/mijnlijst.html?..."`), crawled the songs mentioned on the linked pages, and associated them to the Twitter user account of the original postings.

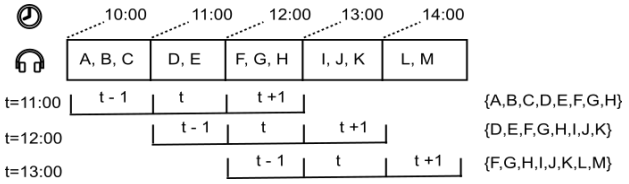


Figure 2: Dynamic Twitter crawling mechanism.

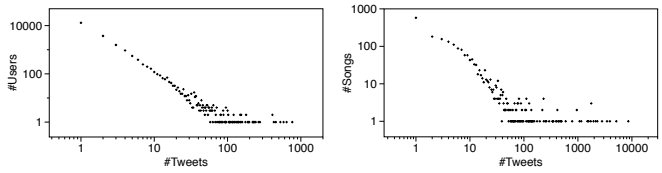
Airing Collection. We also were interested in tweeting behaviour during the airing period, as this would reveal public social and collective engagement with the `Top 2000` content. For this, we employed a dynamic crawling mechanism using the Twitter Streaming API, based on scheduling information of the broadcast. The procedure is visualized in Figure 2: given a set of songs planned for airing in the time interval t, we queried for related microposts in the interval including the hour before $(t-1)$ and the hour after $(t+1)$ the planned airing time. To maximise recall, we configured the Twitter stream listeners with a set of keywords including: 1) the official `#Top2000` hashtag; and 2) the name of the songs and respective artists aired in the targeted slots.

Enrichment With Content-Related Metadata. To better understand the reasons behind the mentioning of songs, we performed sentiment analysis on the microblog post. Common sentiment analysis techniques require prior knowledge of the content's language. While Tweets are often annotated with language metadata, the reported value matches the user's setting, and not the actual language of the content of a tweet. We therefore processed each Twitter micropost with a language detection library[4], after removing song title and artist name from the original content. Sentiment analysis subsequently was performed using SentiStrength[5], configured for usage with the English or Dutch corpora, according to the language of the Tweet. Examples of posts with the strongest detected positive and negative sentiments include: "I love love love #top2000 !!!!", "I just can t stop loving you! MJ top2000", "im angry im gonna miss too many gr8 songs from my fave bands in the #Top2000 because of my job smFh! !!!". "Was terrified of this album cover as a child! #Top2000".

Enrichment With Music-Related Metadata. As a final step of enrichment, we queried the `last.fm` API (`track.getTopTags`) to acquire social tags for the song set. For this, we considered all songs that were mentioned in the voting lists, amounting to 8.441 songs in total. Of these, social tags could be found for 6.341 songs, yielding 371.399 taggings employing a vocabulary of 50.146 unique tags in total. An interesting aspect of social tags is that they give folksonomic descriptions of songs, which are not just limited to genre (e.g. beyond *pop*, *rock* and *cool acid jazz* we also find descriptions like *romantic*, *screaming* and *seaside soundtrack*).

[2]https://en.wikipedia.org/wiki/Top_2000
[3]http://www.nporadio2.nl/nieuws/10587/meer-jonge-luisteraars-dan-ooit-voor-top-2000

[4]Langdetect. https://pypi.python.org/pypi/langdetect
[5]http://sentistrength.wlv.ac.uk/



(a) #tweets per user **(b)** #tweets per song

Figure 3: Distribution of #tweets (a) per user and (b) per song in the airing period.

4. DATASET

Table 1 summarises the main statistics of the resulting dataset, which are further detailed in the next sections.

Voting Collection		Airing Collection	
#lists	1.736	#relevant tweets	712.720
#votes	33.885	#users	482.693
#songs	8.441	#tweets per user	1,48
Avg. votes/user	19	Avg. tweets/song	205,28

Table 1: Descriptive statistics of the dataset.

Voting Collection. The voting collection includes 1:736 casted vote lists, with a total of 33:885 individual votes. The set of voted songs, S_v includes 9:508 distinct songs. 1:900 of them appeared also in the final Top2000 song set (S_f). Thus $|S_v \cap S_f| = 1{:}900$.

Airing Collection. The airing collection includes 4:184:517 tweets. The collection includes a large number of irrelevant posts, retrieved due to the characteristics of the Twitter streaming API. For example, the "Hotel California" filter yields tweets including words "Hotel" and "California" in any order, or at any location within the content.

To increase precision, we applied a set of post-filtering strategies aimed at keeping only tweets containing: 1) the *top2000* keyword; 2) the full title, or 3) the full artist name of a song included in the final `Top 2000` rank. Table 2 reports the statistics of the resulting dataset.

Keywords	Users	Microposts
<Raw Set>	2.369.395	4.184.517
(A) Top2000	21.937	72.333
(B) Full Song Title	321.293	410.565
(C) Full Artist Name	166.643	248.233
(A) or (B) or (C)	482.693	712.720

Table 2: Dataset before and after post-hoc filters.

Figure 3 shows the distribution of number of tweets 1) per users, and 2) per songs mentioned during the airing period. The former conform to a power law distribution; the latter is slightly skewed to smaller values.

The dataset is available for download at the following URL: http://www.wis.ewi.tudelft.nl/MusicWebSci2016. It contains references to the tweets included in the analysis, enriched with information about the language of the micropost, the sentiment expressed in it, and its collection of reference, namely *voting* or *airing*. In the former case, each tweet is associated with the songs included in the casted vote; in the latter case, we indicate the mentioned song(s). For each song, we provide the full title, artist name, and a flag indicating the availability of social metadata on LastFM.

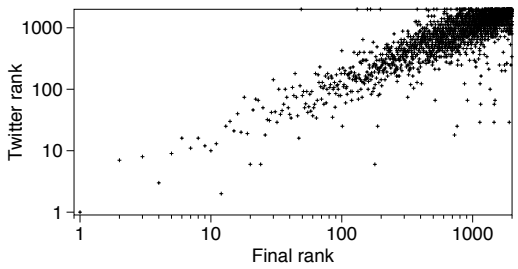


Figure 4: Twitter Rank vs. Final `Top2000` rank.

5. ANALYSIS

This section illustrates two analysis performed on the resulting dataset. We seek answer to two research questions:

RQ1: Is the social media activity related to `Top 2000` votes an accurate reflection of the final preferences nationwide?

RQ2: How is the `Top 2000` airing reflected in the social media stream?

We investigate **RQ1** to show the relevance of the dataset with respect to the studied public voting event. The study of **RQ2** exemplifies the research about the analysis of online music consumption dynamics that can be enabled by our novel dataset.

5.1 Top 2000 rank vs. Twitter rank

To answer **RQ1**, we calculate the correlation between the official `Top 2000` ranking R_f, and a song ranking based on the number votes R_v mentioned in the voting collection. Given that actual number of votes obtained by each song in the final rank is not public, we compare the ranks in R_f and R_v. We used two metrics, namely Spearman's rank correlation, and Kendall's rank correlation. Both are two well accepted measures of non-parametric rank correlations: Spearman's rank correlation bases the calculation on rank deviation while Kendall's rank correlation is based on pairwise agreements between songs. $(R_f; R_v) = 0{:}76$, while $(R_f; R_v) = 0{:}59$, both at $p < 0{:}001$. Figure 4 shows the log-log scaled scatter plot of the R_f and R_v. One could observe a significant correlation between the two ranks, which suggests that votes shared on social media can be a good reflection of the voting preference expressed nationwide.

5.2 Top 2000 Airing Marathon in Twitter

The enrichment process performed on our dataset enables the investigation of **RQ2** from multiple perspective. We focus on four: *Who* mentioned `Top 2000` songs during the airing marathon; *What* was the sentiment expressed on social media; *When* the mention took place, w.r.t. the actual airing time; and *Where* the mentions where expressed on the Dutch territory.

Dataset Enrichment With User-Related Metadata. To provide a demographic characterisation of the song mentioning behaviour during the airing of `Top 2000`, we processed user profiles to infer information about the *area of residence*, *gender*, *age*, of users in the dataset.

Users' location is calculated using microposts' geolocation tags, Twitter user settings, and the SocialGlass [7] platform. In absence of geo-located microposts, we checked the

user profile for user-provided location information. 61,9% of the users in the dataset filled in the location field in their user settings. To normalise the geographic information, we converted all locations into coordinates using the Geonames database[6]. 49,1% of location names could be matched with geographic coordinates. Finally we used SocialGlass to infer the location of users according to the history of geo-located microblog activity of users. The combined outcomes of our integrated geocoding process resulted in a location information for 50,1% of users.

Age and gender metadata were collected from SocialGlass, using the method described in [7].

Metric	Gender		Age			Country	
	Male	Female	Young	Mid-aged	Older	NL	Non-NL
# Users	10.656	7.031	3.460	3.722	1.951	6.405	553
# Posts	32.226	19.785	7.438	8.216	4.840	18.795	819
Avg.	3,02	2,81	2,15	2,21	2,48	2,93	1,48

Table 3: Users' activity by gender, age, and country.

User Analysis. We group users in the following age groups: 1) *Young* – 15 to 30; 2) *Middle-aged*: 31 to 45; and 3) *Older*: above 45 years of age. In this age-based user grouping we consider younger ages than traditional literature in social and physiological science (e.g. [11, 5]) as the use of social media is more familiar to the younger generations [8]. Results are summarised in Table 3. Young and Mid-aged users have comparable presence and activity, whereas Older users are in relatively smaller number. The result indicates an evenly distributed interest across age groups; male users, on the other hand, are prevalent in the dataset, also in terms of number of posts.

The majority of users appear to reside in the Netherlands, an expected result. An analysis on the language of the considered microposts (Table 4), however, reveals the presence of relatively more non-Dutch speaker users, although Dutch speaking users are still the majority. Such an observations hints a multiculturalism property of residents (as music consumers) in the Netherlands.

lang	EN	NL	AF	Other
Top2000	6.200	54.696	9.191	2.246
Entire	547.677	64.096	15.492	85.455

Table 4: Statistics of content language.

Content Analysis. Figure 5 shows the result of sentiment analysis on the microposts including "Top200" keyword. Overall, the number of non-negative and non-positive sentiments constitute the largest group, which may mean most of the microposts are factual or informational. The number of tweets with negative sentiments is however lower than the number of positive sentiments. This is not surprising, considering that `Top 2000` is an entertainment event in which participation is optional. As a final remark, we observe that the number of microposts having no negative sentiment appears as being the largest group among microposts in all ranges. This observation aligns with the general sentiment distribution of social media users towards large-scale events [13].

[6]http://www.geonames.org/

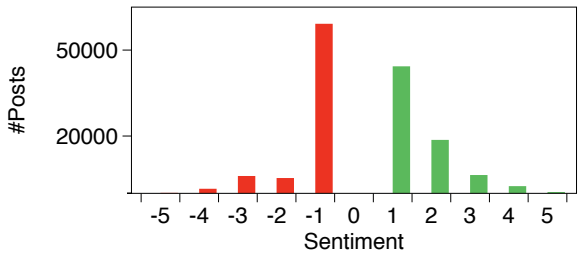


Figure 5: Sentiment analysis. From -5 to 5 the sentiment varies from highly negative to highly positive.

Temporal Analysis. The focus is on microposts containing the keyword "Top2000", and on the distribution of song mentions in the 159 hours of the airing marathon. For each airing hour i, we consider all songs aired at this hour, denoted as $Songs_i$. We then aggregate, in every hour j, the number of micropost mentions to any song in $Songs_i$, denoted by $M_{Songs_i} = \{m_{ij} | 1 \leq j \leq 159\}$, where m_{ij} is the aggregated #mentions of songs in $Songs_i$ in hour j. To show the relationship that exists between the airing of songs and the mentioning behaviour of users, we construct a matrix, depicted in Figure 6a. Each entry $(i; j)$ shows the aggregated (and normalised) number of mentions at hour j of all songs aired at hour i. Higher intensity in the secondary diagonal clearly indicates that songs are mentioned much more in Twitter at the hour they are aired. However, songs were mentioned also at different times. This is shown in Figure 6c: the airing hour of songs corresponds with a clear peak, although mentions are distributed, with lower intensity, across the whole period. Figure 6d shows another interesting angle of analysis, by depicting, for each airing hour, the amount of mentions for songs aired during that hour. We can observe how mentioning patterns greatly vary during the event; interestingly, the day when the most popular songs are aired (Dec. 31st) is the one with, on average, less mentions in Social Media. Further analysis on this result are left to future work.

Spatial Analysis. Figure 6b summarises the spatial distribution of the mentioning microposts published during the airing period. The map clearly shows how the event involved the whole country, with more intense activities around the main cities (e.g. Amsterdam, Rotterdam, Utrecht, etc.).

6. CONCLUSIONS

This paper contributes a dataset of microblogging activity surrounding the `Top 2000`, a significant Dutch yearly music ranking and broadcasting event. We show the social media activity to provide a good reflection of the national voting preferences. Finally, we discuss a study of the mention behaviour of users during the airing of the `Top 2000`, showing examples of data-oriented studies of the public response to, and public significance of, the aired songs.

This work shows how future work can develop in multiple directions. Our ultimate goal is to understand how Web data can be used to reveal latent social and collective perspectives on music. In this respect, we will investigate how user demographic properties (e.g. location) can be used as pivoting dimensions for further analysis on voting and mention behaviour.