



# MuSe: The Musical Sentiment Dataset

**CHRISTOPHER AKIKI** 

**MANUEL BURGHARDT** 

*\*Author affiliations can be found in the back matter of this article*

DATA PAPER

]u[ubiquity press

## ABSTRACT

The MuSe (Music Sentiment) dataset contains sentiment information for 90,001 songs. We computed scores for the affective dimensions of valence, dominance, and arousal, based on the user-generated tags that are available for each song via Last.fm. In addition, we provide artist, title and genre metadata, and a MusicBrainz ID and a Spotify ID, which allow researchers to extend the dataset with further metadata.

## CORRESPONDING AUTHOR:

**Christopher Akiki**

Text Mining and Retrieval  
Group, Leipzig University,  
Germany

[christopher.akiki@uni-leipzig.de](mailto:christopher.akiki@uni-leipzig.de)

---

## KEYWORDS:

music information retrieval;  
music emotion recognition;  
music sentiment; music  
dataset; sentiment analysis

## TO CITE THIS ARTICLE:

Akiki, C., & Burghardt, M.  
(2021). MuSe: The Musical  
Sentiment Dataset. *Journal  
of Open Humanities Data*, 7:  
10, pp. 1–6. DOI: [https://doi.  
org/10.5334/johd.33](https://doi.org/10.5334/johd.33)

## (1) OVERVIEW

### REPOSITORY LOCATION

DOI: <https://doi.org/10.34740/kaggle/dsv/2250730>

URL: <https://www.kaggle.com/cakiki/muse-the-musical-sentiment-dataset>.

### CONTEXT

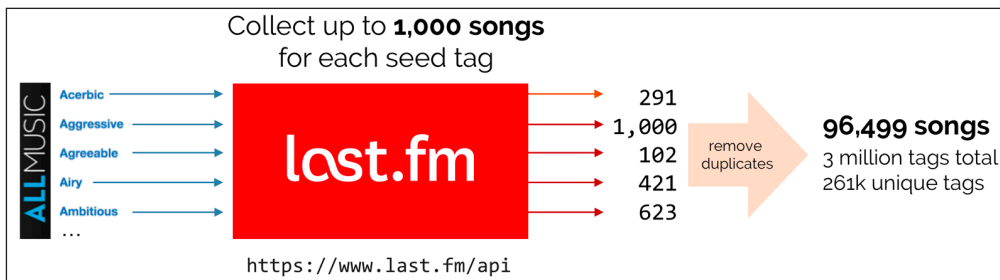
The study of sentiments and emotions, and in particular their influence on human cognitive processes, has long been part of the research agenda in psychology and cognitive studies. Increasingly, sentiment analysis methods are also being applied in the digital humanities, where they are used to study the interplay between emotions and sociocultural artifacts such as literary texts (Kim & Klinger, 2019). With the dataset presented in this paper, we hope to provide a useful resource that will stimulate further studies on sentiment analysis in the field of musicology (see also Akiki & Burghardt, 2020).

## (2) METHOD

This section provides an overview of the basic steps and resources that were involved in creating the dataset; for more details on each of the steps, see Akiki and Burghardt (2020).

### STEPS

1. *Seeding*: In this step, we used 279 mood labels from AllMusic (<https://www.allmusic.com/moods>, last scraped Sep. 2019) as seeds to collect song objects from the Last.fm API (<https://www.last.fm/api>). The AllMusic mood labels that were used are documented in a “seeds” column in the dataset.
2. *Expansion*: For each of the 279 seed moods, we collected up to 1,000 songs, which is currently the official limit of the Last.fm API. As we did not retrieve the maximum of 1,000 songs for each of the seed labels, we ended up with a total of 96,499 songs (see [Figure 1](#)).



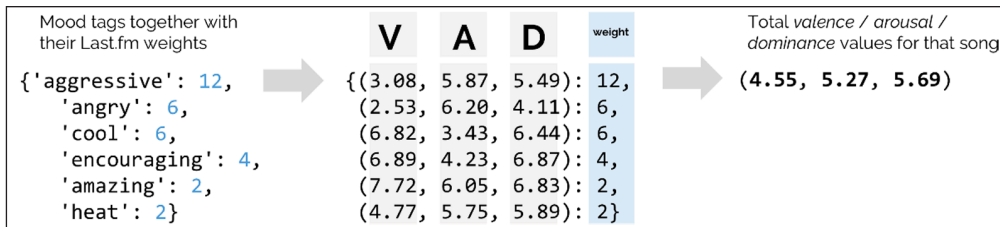
**Figure 1** The sampling process is guided by 279 seed mood labels from AllMusic, which were used to retrieve songs with basic mood labels via the Last.fm API.

3. *Filtering*: Next, we filter the more than 261k unique user-generated tags by using the WordNet-Affect list (Strapparava & Valitutti 2004). This step is inspired by related work from Hu, Downie, and Ehmann (2009) and Delbouys, Hennequin, Piccoli, Royo-Letelier, and Moussallam (2018), and leaves us with a list of songs that contain a least one mood-related tag.
4. *Mapping*: Now, we map the identified mood tags to Russell’s (1980) “circumplex model” of *valence* and *arousal* and extend it by the third dimension of *dominance*, as suggested by Scherer (2004). The mapping is achieved by using the wordlist by Warriner, Kuperman, and Brysbaert (2013), which contains crowdsourced values for *valence*, *arousal* and *dominance* (V-A-D) for a total of 13,915 lemmas. [Table 1](#) provides some statistics for the distribution of the V-A-D tags within our dataset.

As songs often have multiple mood tags, we calculate the weighted average for V-A-D (see [Figure 2](#)). The weights for each tag are derived from Last.fm, where higher values indicate an increased relevance of a tag.

	VALENCE TAGS	AROUSAL TAGS	DOMINANCE TAGS
<b>Count</b>	90,001	90,001	90,001
<b>Mean value</b>	5.45	4.32	5.25
<b>Min value</b>	0.23	0.11	0.23
<b>Max value</b>	8.48	7.27	7.44
<b>Std.</b>	1.55	1.15	1.17

**Table 1** Basic statistics for all V-A-D tags in the dataset.



**Figure 2** Example for a song with its mood tags and tag weights (left), the individual V-A-D scores for each tag (middle) and the weighted average for the three dimensions (right). Image taken from Akiki and Burghardt (2020).

Since some of these songs do not contain a tag that matches the V-A-D wordlist by Warriner, Kuperman, and Brysbaert (2013), a total of 90,408 songs remain after the mapping step.

- 5. Metadata:** As a final step, we add information that allows researchers to extend the dataset with further metadata. From the Last.fm API, we collect the MusicBrainz ID (*mbid*; [https://musicbrainz.org/doc/Developer\\_Resources](https://musicbrainz.org/doc/Developer_Resources)) for each song for which it is available (this is the case for about two-thirds of the dataset). The *mbid* also allows us to remove duplicates, which occur when the artist name or song title is spelled differently. After removing 407 duplicates from 61,624 songs with a *mbid*, we are left with a total of 90,001 songs in our dataset. There may be a few more duplicates in the dataset among the songs that do not come with a *mbid*. To showcase a potential enhancement of the dataset, we also collected the Spotify ID for a total of 61,630 songs, which enables researchers to add further metadata from the Spotify API (<https://developer.spotify.com/documentation/web-api>). More details on how we added the Spotify ID are described in the reuse section of this paper.

### (3) DATASET DESCRIPTION

The dataset contains the following types of information for each song object (see **Table 2**) and is available for download in CSV format.<sup>1</sup>

ATTRIBUTE	DESCRIPTION	EXAMPLE VALUE
<i>lastfm_url</i>	Last.fm page of the song	<a href="https://www.last.fm/music/eminem/_/%2527till%2bi%2bcollapse">https://www.last.fm/music/eminem/_/%2527till%2bi%2bcollapse</a>
<i>track</i>	Song title	'Till I Collapse'
<i>artist</i>	Artist name	Eminem
<i>seeds</i>	The initial keyword(s) that seeded the scraping of this song	['aggressive']
<i>number_of_emotion_tags</i>	Number of words that contributed to the emotion score of the song	6
<i>valence_tags</i>	Pleasantness dimension of the song	4.55
<i>arousal_tags</i>	Intensity dimension of the song	5.27
<i>dominance_tags</i>	Control dimension of the song	5.69
<i>mbid</i>	MusicBrainz Identifier of the song	cab93def-26c5-4fb0-bedd-26ec4c1619e1
<i>spotify_id</i>	Spotify Identifier of the song	4xkOaSrKexMciUUogZKVTS
<i>genre</i>	Genre of the song	rap

**Table 2** Information available in the MuSe dataset.

<sup>1</sup> Please note that the current dataset is in version 3, as it contains slightly different metadata than originally described in Akiki and Burghardt (2020). Further adjustments were made following the peer review process of this paper.

*A note on genre:* The user-generated Last.fm tags not only contain emotion tags, but also a vast amount of genre tags. We extract these genre tags by filtering the weighted list of tags against a hardcoded list of musical genres, which essentially leaves us with a list of weighted genres describing each song.<sup>2</sup> We then assume the genre label with the highest weight to be the most likely representative of a given song's genre and include that in the dataset. Using this method, we were able to label 76,321 songs with genre information.

*A note on missing release years:* Information about the release year of a song is unfortunately not available via the Last.fm API.

## OBJECT NAME

- muse\_v3.csv (17.36 MB)

## FORMAT NAMES AND VERSIONS

CSV

## CREATION DATES

from 2020-09-01 to 2021-05-20

## DATASET CREATORS

- Christopher Akiki (Text Mining and Retrieval Group, Leipzig University): Conceptualization, Data Curation, Formal Analysis, Methodology

## LANGUAGE

English

## LICENSE

CC BY 4.0

## REPOSITORY NAME

MuSe: The Musical Sentiment Dataset – 90K Songs in Three-dimensional Sentiment Space

## PUBLICATION DATE

2021-05-20

## (4) REUSE POTENTIAL

With our current MuSe dataset, we provide a resource that enables different kinds of research questions that take into account the relationship between the V-A-D dimensions and other metadata, such as *artist*, *title*, and *genre*. As the mood tags themselves cannot be included in the dataset for copyright reasons, we provide a Jupyter notebook via the Kaggle repository that demonstrates how to fetch the tags of a given song from the Last.fm API (<https://www.kaggle.com/cakiki/muse-dataset-using-the-last-fm-api>).

To illustrate the reuse potential of the data set in terms of extensibility, we also provide the Spotify ID whenever we were able to find one in an unambiguous way for the songs in our collection (see Akiki & Burghardt, 2020). In the end, we were able to track down a Spotify ID for a total of 61,630 songs. Via the Spotify ID, researchers may append any additional information to the dataset that is available via the Spotify API, for instance:

- *further metadata*: release date, popularity, available markets, etc.
- *low-level audio* features: bars, beats, sections, duration, etc.
- *mid-level audio features*: acousticness, danceability, tempo, energy, valence, etc.

---

<sup>2</sup> The genre labels were taken from <https://everynoise.com/everynoise1d.cgi?scope=all> (last accessed on May 15th, 2021).

With this additional information, research questions such as the following could be investigated: Is there a correlation between a song's popularity rank and its sentiment? Are there genre-specific effects, i.e., does negative sentiment help the popularity of songs in certain genres (e.g., blues or black metal) but tend to negatively affect the popularity of songs in other genres (e.g., pop and dance)?

In the Kaggle repository of the dataset, we provide another Jupyter notebook (<https://www.kaggle.com/cakiki/muse-dataset-using-the-spotify-api>) that demonstrates how to enrich the dataset with audio features using various endpoints of the Spotify API (as showcased in Akiki & Burghardt, 2020). Another way to extend the dataset with further metadata is provided by means of the MusicBrainz ID, which we gathered directly from the Last.fm API for about two-thirds of the songs. MusicBrainz provides additional information for each song, for instance, the respective cover art, which allows for future studies on the relationship of musical sentiment and the cover art design. Furthermore, artist and title information may be used to add lyrics information to analyze the relation of lyrics and musical sentiment. In addition to Spotify and MusicBrainz, other sources of metadata, most notably Discogs (<https://www.discogs.com/>), might be added to the dataset.

We believe that MuSe will be a good starting point for musical sentiment data that can be extended in several directions. All in all, we hope the MuSe dataset will help to advance the field of computational musicology and thus provide an incentive for more quantitative studies on the function and role of emotions in music.

## ACKNOWLEDGEMENTS

We are grateful to Last.fm for allowing us to share selected metadata for more than 90k songs as part of this dataset.

## COMPETING INTERESTS

The authors have no competing interests to declare.


## AUTHOR CONTRIBUTIONS

*Christopher Akiki*: Conceptualization, Data Curation, Formal Analysis, Methodology.

*Manuel Burghardt*: Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

## AUTHOR AFFILIATIONS

**Christopher Akiki**  [orcid.org/0000-0002-1634-5068](https://orcid.org/0000-0002-1634-5068)  
Text Mining and Retrieval Group, Leipzig University, Germany

**Manuel Burghardt**  [orcid.org/0000-0003-1354-9089](https://orcid.org/0000-0003-1354-9089)  
Computational Humanities Group, Leipzig University, Germany

## REFERENCES

- Akiki, C., & Burghardt, M.** (2020). Toward a Musical Sentiment (MuSe) Dataset for Affective Distant Hearing. *Proceedings of the 1st Workshop on Computational Humanities Research (CHR)*, 225–235. <http://ceur-ws.org/Vol-2723/short26.pdf>
- Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., & Moussallam, M.** (2018). Music Mood Detection Based on Audio and Lyrics with Deep Neural Net. *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 370–375.
- Hu, X., Downie, J. S., & Ehmann, A. F.** (2009). Lyric Text Mining in Music Mood Classification. *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, 411–416.
- Kim, E., & Klinger, R.** (2019). A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *Zeitschrift für digitale Geisteswissenschaft*, 4. DOI: [https://doi.org/10.17175/2019\\_008](https://doi.org/10.17175/2019_008)
- Russell, J.** (1980). A Circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. DOI: <https://doi.org/10.1037/h0077714>

- Scherer, K. R.** (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *Journal of New Music Research*, 33(3), 239–251. DOI: <https://doi.org/10.1080/0929821042000317822>
- Strapparava, C., & Valitutti, A.** (2004). WordNet Affect: An Affective Extension of Word-Net. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, 1083–1086.
- Warriner, A. B., Kuperman, V., & Brysbaert, M.** (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4), 1191–1207. DOI: <https://doi.org/10.3758/s13428-012-0314-x>

Akiki and Burghardt  
*Journal of Open  
Humanities Data*  
DOI: 10.5334/johd.33

TO CITE THIS ARTICLE:

Akiki, C., & Burghardt, M. (2021). MuSe: The Musical Sentiment Dataset. *Journal of Open Humanities Data*, 7: 10, pp. 1–6. DOI: <https://doi.org/10.5334/johd.33>

Published: 07 July 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press.