P4KxSpotify: A Dataset of Pitchfork Music Reviews and Spotify Musical Features

Anthony T. Pinter, Jacob M. Paul, Jessie Smith, Jed R. Brubaker
Department of Information Science
University of Colorado Boulder
Boulder, CO 80309
{anthony.pinter; jacob.paul; jessie.smith-1; jed.brubaker}@colorado.edu

Abstract

Algorithmically driven curation and recommendation systems like those employed by Spotify have become more ubiquitous for surfacing content that people might want hear. However, expert reviews continue to have a measurable impact on what people choose to listen to and the subsequent commercial success and cultural staying power of those artists. One such site, Pitchfork, is particularly known in the music community for its ability to catapult an artist to stardom based on the review that an album receives. In this paper, we present P4KxSpotify: a dataset of Pitchfork album reviews with the corresponding Spotify audio features for those albums. We describe our data collection and dataset creation process, including the ethics of such a collection. We present basic information and descriptive statistics about the dataset. Finally, we offer several possible avenues for research that might utilize this new dataset.

Introduction

With the proliferation of online review sites, what was once left to experts is now a common practice that anyone can take part in—leaving restaurant reviews on Yelp,[1] reviews of city attractions on TripAdvisor,[2] and offering critiques of artistic expression such as movies and music on MetaCritic.[3] However, this shift has not rendered the expert review obsolete. In music, sites like Pitchfork[4] curate album reviews from expert music journalists, and those reviews are often held up as the gold standard in reviewing for music. When faced with numerous dissenting reviews from many users on collaborative reviewing sites, one can turn to Pitchfork and expect to find the most trusted voice of what music is the best music, right now.

Other ways of finding new music that are more tailored to one's individual tastes have emerged through services such as Pandora[5] or Spotify. Spotify is one of the largest music streaming services in the world, with a library of over 50 million tracks and 248 million users across 79 countries.[6] Spotify also makes a wealth of data about music available to developers via its API,[7] including the ability to examine the audio features that describe individual tracks. Spotify uses these features internally to help generate playlists that are unique to each user's listening habits, such as Discover Weekly[8] and Release Radar.[9] Meanwhile, other applications for these audio features have emerged, like Obscurify, an application that compares an individual's listening habits to the general populace[10] and Soloman Goldfarb's project to find songs in an artist's catalog that might be enjoyable based on a different song one already likes (i.e., if a person likes "Seven Nation Army" by The White Stripes, they will probably also enjoy "Icky Thump" and "Rag and Bone") (**?**). Instead of relying on subjective reviews (however unbiased they claim to be) to find music, users can now look to computationally derived attributes and recommendation algorithms to find new music to consume.

However, neither method of finding new music is perfect. Just because an album is rated highly by a music journalist does not mean that one will enjoy that album. Similarly, many users of Spotify have shared anecdotes of being recommended music they did not like. Yet, no dataset to our knowledge exists that combines the subjectivity of expert reviews with the objectivity yielded by computationally derived audio features that describe music.

To remedy this, we present a dataset we call **P4KxSpotify**. P4KxSpotify consists of Pitchfork reviews along with the audio features of those albums from Spotify. We scraped Pitchfork, a review site that has been publishing reviews since the late-1990's and currently has over 20,000 reviews in its catalog spanning nine genres. We then used Spotify's API to retrieve 10 audio features that describe each album. This process yielded a dataset that enables researchers to ask new questions about how to best evaluate and recommend cultural artifacts to consumers,

[1] http://yelp.com
[2] https://www.tripadvisor.com
[3] https://www.metacritic.com
[4] https://pitchfork.com
[5] https://www.pandora.com
[6] https://newsroom.spotify.com/company-info/
[7] https://developer.spotify.com/documentation/web-api/
[8] https://www.spotify.com/us/discoverweekly/
[9] https://artists.spotify.com/blog/say-hello-to-release-radar
[10] https://obscurifymusic.com/

among other applications.

The remainder of the paper is structured as follows: we begin by discussing other initiatives aimed at collecting large corpora of review data to better illustrate the novelty and utility of our dataset. Next, we report in-depth on data collection and dataset composition as well as the ethics of the collection process. Then, we describe the dataset, including its schema and basic descriptive statistics. Lastly, we discuss uses for this dataset and offer suggestions for future research utilizing this dataset.

Existing Datasets of Reviews

As the number of websites that collect and aggregate reviews (expert or user) continues to grow, there have been numerous datasets collected of these reviews and associated metadata. For example, researchers at UCSD led by Jianmo Ni have collected Amazon reviews iteratively over the past 7 years, yielding a dataset of Amazon products that include 233.1 million reviews, product data, and links.[11] Often, these datasets are created with sentiment analysis or other machine learning techniques in mind as seen in the work by Ni, et al. (**?**), Maas, et al. (**?**), and Pang et al. (**?**).

Here, we combine subjective measures of quality (reviewer's scores) with objective measures of the artefact being reviewed (computationally derived descriptors of the music)— an approach that has not been used in prior work.

Data Collection

Our dataset is composed of data from two sources: the review section of Pitchfork's website and the Spotify public-facing API. We collected this data in a two-step process, first scraping Pitchfork and then using the results of that scrape to pull data from the Spotify API.

Pitchfork Data Collection

We completed the Pitchfork portion of data collection in three steps. First, we used an implementation of Selenium[12] to scrape the URLs of each album review from the Pitchfork "Reviews" page.[13] Next, we used a second implementation of Selenium along with URLlib.request[14] to open each URL from the previous step and save the raw HTML of the page. Last, we used BeautifulSoup[15] to parse the data of interest from each saved HTML file. Our scrape took place in two parts: the first part occurred in early June 2019, while the second part occurred in early January 2020 to capture the remainder of 2019 in the dataset. In total, these two scrapes yielded 22,060 distinct reviews, the total number of reviews on the Pitchfork website that were published in 2019 or earlier.

[11] https://nijianmo.github.io/amazon/index.html

[12] https://selenium.dev

[13] https://pitchfork.com/reviews/albums/

[14] https://docs.python.org/3/library/urllib.request.html

[15] https://beautiful-soup-4.readthedocs.io/en/latest/

Spotify Data Collection

We built a Python script that collected data from the Spotify API in three steps: album URI (Uniform Resource Identifier) retrieval, track URI retrieval, and track feature retrieval.

We began by searching for individual albums from the Pitchfork data to obtain album URIs from Spotify. The album name was used as the primary search term from which the API would then return possible matches. Artist name and release year were used as secondary terms to computationally verify which returned search result was the closest match.

The first 25 album results were computationally examined to identify a matching result based on a set of criteria. First, the album name had to be at least 70% similar to the queried album name. We used a Python library called difflib[16] to compute the similarity score between these two values using a native adaptation of the gestalt pattern matching algorithm developed by Ratcliff and Obershelp (**?**). If the album names were similar according to the algorithm, we checked for matches between three datapoints: artist name, album name, and release year. If album name and artist name matched, or if album name and release year matched, the album URI was accepted. If only the album name matched, we manually verified the album and included it in the case of a match.

Next, the track URIs for the tracks on the matched albums were retrieved by using the get_album_tracks endpoint provided by the Spotify API. The album URI was passed to the API and the track list, which contained the track URIs, was returned. All of the track URIs were stored outside of the original dataset, so that they could be grouped when being passed into the track features API endpoint. The track features endpoint can retrieve track features for up to 100 songs per API call. So, the list of track URIs was split into groups of 100 and then passed to the track features endpoint. These track features were then parsed out of the API response and merged with the track and album URIs to create a tabular track dataset that included every track, its corresponding URIs, and its track features. Any tracks where the track features were unavailable were eliminated from the dataset.

Using the dataset of track features, we calculated the mean value of each audio feature for the album as a whole. Finally, this aggregated dataset was joined to the original dataset of Pitchfork review data to create our final dataset of both the data from Pitchfork as well as the mean Spotify track feature scores for each album.

Our final dataset was comprised of 18,403 entries, representing a loss of 3,657 albums (or 16.6%). This loss was the result of: (1) our decision to skip any album where the artist was "Various Artists" because this was often a source of mismatch between Spotify's naming conventions and Pitchfork's naming conventions (i.e., Spotify would label the album under the "Various Artists" tag, while Pitchfork used a different name); or (2) the album not being present in Spotify's US library.

[16] https://docs.python.org/3/library/difflib.html

Ethics of Data Collection

As we constituted our dataset through web scraping and API pulls, we carefully considered the ethical implications of collecting such data. While both Conde Nast[17] (Pitchfork's parent company) and Spotify[18] have provisions in their terms of service and/or user agreement prohibiting scraping, recent work has sought to draw a distinction between violating a service's terms of service from a legal standpoint versus an ethical standpoint (**?**). As suggested by Fiesler et al., limiting data collection solely to instances where terms of service allows it "suggests that violating TOS is (a) inherently unethical; and (b) the only reason that data collection could be unethical."

Instead, Fiesler et al. suggest that the ethicality of creating a dataset should instead be judged using heuristics that align closely with the Belmont Report, which guides human subjects research (**?**; **?**). With this in mind, we argue that the dataset we present here was collected in an ethical manner for the following reasons that are closely tied to the concepts of *beneficence* and *justice* found in the Belmont Report (**?**):

- We were careful to collect data in a manner that did not put undue load upon a service's servers.
- The data collected from both services is publicly accessible, and consists of public (or semi-public) individuals and groups.
- Constituting the dataset offers unique opportunities not only for a variety of research initiatives (as detailed later in this paper), but also potentially to the two services themselves. For Spotify, it offers an opportunity to understand what albums are not available for US consumption, but are culturally important according to a major review site; for Pitchfork, it offers an opportunity to understand how implicit biases towards certain types of music might manifest in their ratings.

Next, we discuss where to find the dataset and provide basic descriptive statistics and visualizations of the dataset.

Dataset Description

In this section, we discuss the publication of the dataset and the scripts used to constitute the dataset, and present basic descriptive statistics for key columns of the dataset.

Publication of Dataset and Supplementary Material

The P4KxSpotify dataset is published on Zenodo at https://zenodo.org/record/3603330#.XheASC3MzOQ and can be referenced with the DOI 10.5281/zenodo.3603330. All other supporting material, including the scripts used to constitute the dataset, raw HTML, and album art, can be found at https://github.com/cuinfoscience/objectively-reviewed. The code used for basic descriptive statistics below was written in Python 3 in a Jupyter Notebook and can be found in the same GitHub repository.

[17]https://www.condenast.com/user-agreement/

[18]https://developer.spotify.com/terms/#iv

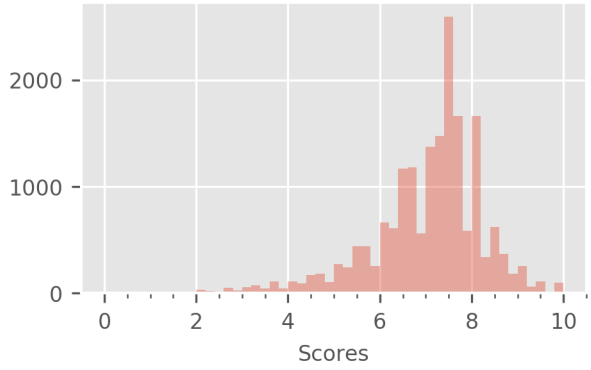

Figure 1: Score distribution of album reviews in the dataset.

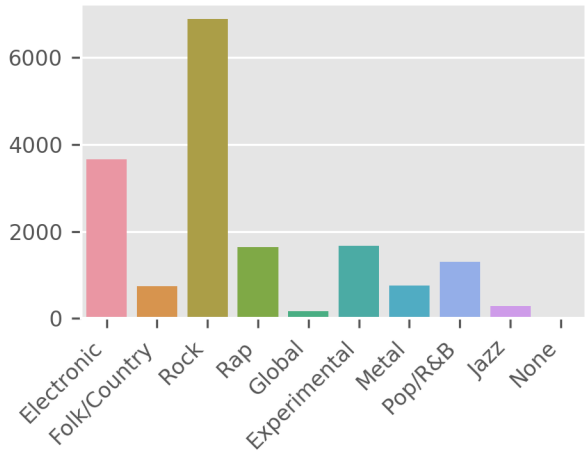

Figure 2: Genre counts of albums in the dataset.

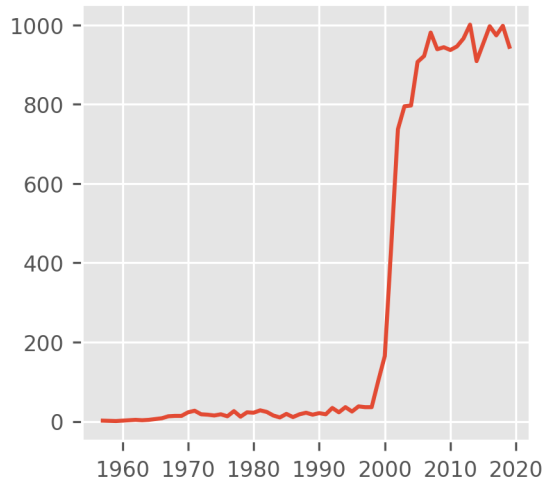

Figure 3: Count of albums reviewed by the year of release.

Table 1: P4KxSpotify dataset column names and descriptions

Column Name	Column Description
artist	The name of the artist who created the album being reviewed.
album	The name of the album being reviewed.
recordlabel	The name of the record label(s) who published the album.
releaseyear	The year that the album was released.
score	The score given to the album by the reviewer on a scale of 0.0 to 10.0.
reviewauthor	The name of the author who reviewed the album.
genre	The genre assigned to the album by Pitchfork.
reviewdate	The date that the review was published.
key	The estimated overall musical key of the track. Integers map to pitches using standard Pitch Class notation (e.g., 0 = C, 2 = D, and so on).
acousticness	A confidence measure from 0.0 to 1.0 of whether an album is acoustic. 1.0 represents high confidence that the album is acoustic.
danceability	How suitable an album is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 1.0 is most danceable.
energy	A perceptual measure of intensity and activity, from 0.0 to 1.0, where 1.0 represents high energy. Metal is often high energy.
instrumentalness	Predicts whether an album contains no vocals, from 0.0 to 1.0. The closer to 1.0, the more likely the album contains no vocals.
liveness	Detects the presence of an audience, from 0.0 to 1.0. Scores greater than 0.8 indicate a strong likelihood the album is live.
loudness	The overall loudness of the album in decibels (dB).
speechiness	Measures the presence of spoken words in an album on a scale from 0.0 to 1.0. Scores higher than 0.66 indicate an album made entirely of spoken words, while scores below 0.33 indicate music and other non-speech-like elements.
valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by an album, where values closer to 1.0 indicate more positive sounds.
tempo	The overall estimated tempo of an album in beats per minute (BPM).

The dataset is released under the *Creative Commons - Attribution - 4.0 International* (CC-NY 4.0) license.[19] The code used to constitute the dataset is released under the GNU General Public License version 3.[20]

Basic Descriptive Statistics

The P4KxSpotify dataset consists of 18,403 records described by 18 columns. The dataset covers a period of publication from 1957 through the end of 2019. In Table 1, we present the columns with a brief description of what each column represents. The Pitchfork descriptions are intuited from a standard review page; Spotify metrics are described in the API documentation.[21] Next, we present basic statistics, and visualizations when appropriate, for key columns in the dataset, including score, genre, reviewauthor, and releaseyear from Pitchfork, and the ten audio features retrieved for each album from the Spotify API.

Score Pitchfork reviews are scored on a quantitative scale from 0.0 to 10.0, where 10.0 is the highest attainable score. In Figure 1, we present a distribution of scores of the 18,403 reviews captured in our dataset. The average review received a score of 7.03 with a standard deviation of 1.25.

Genre Pitchfork categorizes the albums it reviews into nine distinct genres. Albums can receive more than one genre designation, however, in this dataset, we only included the primary (i.e., first listed) genre for each album. We present a distribution of the genres in Figure 2.

There are 11 albums whose genres are categorized as 'None' in the dataset. These album reviews were all published in the second half of 2019 and did not include a genre category on the review page on Pitchfork at the time of scraping.

Review Author There are 564 review authors represented in the dataset. The review author column contains the name of the author responsible for each review, which becomes part of that author's unique URL on Pitchfork's site. The average reviewer has 32.63 reviews, with a standard deviation of 77.3 reviews. The reviewer with the greatest number of reviews in the dataset is Ian Cohen, with 749 reviews. There are 45 authors with more than 100 reviews in the dataset; these authors' reviews constitute 10,720 of the 18,403 reviews in the dataset, or 58.2% of the total number of reviews. In Figure 4 we present a distribution of the count of reviews.

Release Year In Figure 3, we present the number of albums reviewed by the year of publication (i.e., when the album was originally released). Pitchfork began reviewing al-

[19] http://creativecommons.org/licenses/by/4.0/

[20] https://www.gnu.org/licenses/gpl-3.0.en.html

[21] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

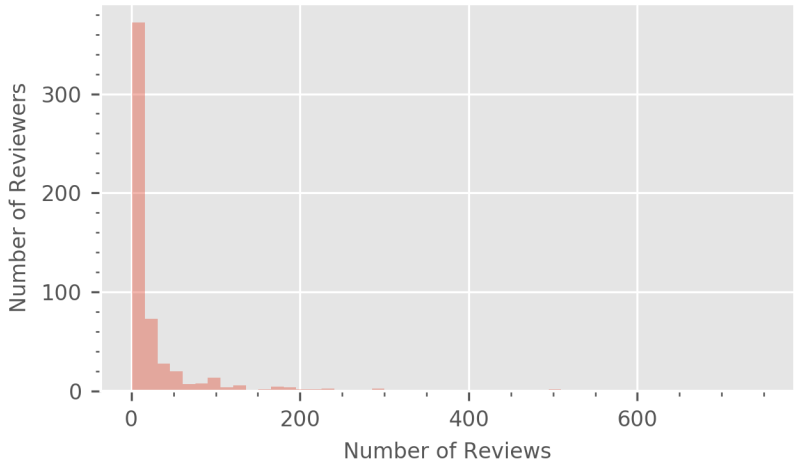

Figure 4: Distribution of the number of authors by reviews represented in the dataset.

bums in 1999, but the earliest review in this dataset is from April 1, 2002 because of the record loss described in our description of data collection. The earliest reviewed album by album publication date is 1957, likely from a series of retrospective album reviews, which Pitchfork generally runs on Sundays.

Audio Features In Figure 5, we present visualizations for comparison to the published Spotify visualizations for the same features in the Spotify API documentation.[22] The visualizations of the P4KxSpotify dataset closely mirror those present in the Spotify API documentation.

In the next section, we discuss potential use cases for this dataset.

Research Opportunities

The P4KxSpotify dataset combines data from computational sources and data from subjective sources into one dataset. Thus, it might be useful to researchers focused on using one form of data to improve the other (i.e., using subjective data to better contextualize computational data, or using computational data to better categorize subjective data). In this section, we offer three such possibilities for research that this dataset might be used in. This section is not meant to be an exhaustive list of the possible uses for the dataset; instead, it is meant to illustrate a variety of uses that this dataset might have to a wide range of researchers and applications.

Investigating Bias in Reviews

Pitchfork as a journalistic review entity has been subject to numerous accusations of bias in their reviews. With their tagline, "The Most Trusted Voice in Music," and an observable effect on the commercial success of an album based on the score given in reviewing that album, Pitchfork's reviews have a significant cultural and economic impact on how listeners might perceive released music.

[22]https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

Bloggers and other non-scholarly sources have attempted to identify bias in Pitchfork's reviews (e.g., (**?**)). The most recent attempt at this was Grantham's 2015 analysis of individual author trends on Pitchfork, where he grouped reviewers into two categories based on review data – conservative (averaging 6.0-6.5 scores) and liberal (averaging 7.0-7.5 scores) (**?**). Grantham's analysis demonstrates that reviewers have scoring tendencies, which in turn can have significant impact for artists. Others have also taken this issue up, with Briskin finding that score was less important than receiving one of Pitchfork's coveted accolades (**?**).

Research focused on reviewing practices in other contexts has found that bias does exist and has enumerated factors that contribute to that bias (e.g., (**?**; **?**; **?**)). Our dataset offers the opportunity to evaluate Pitchfork's reviews in a more systematic way that would contribute to the body of research concerned with bias in reviews. Our methodology reported here also offers a guide for composing datasets from other review sites that could ultimately yield an understanding of how bias appears in reviews.

Applications in Recommendation Systems

As online services provide access to ever-larger music collections, music recommendations have become increasingly important. Music recommender systems are needed, not only for song recommendation but also for playlist generation and Music Information Retrieval (MIR) (**?**). The three main techniques used for music recommendation are content-, metadata-, and hybrid-based models (**?**; **?**). Content-based models consider the musical properties of songs when creating recommendations. However, content-based music recommendations are often inaccurate due to the assumption that a user's favorite pieces are acoustically similar, which is not always true (**?**). In contrast, metadata-based models consider non-musical data from songs such as artist name and genre. These include collaborative-based models, which consider how other users have rated pieces when creating recommendations (**?**). How-

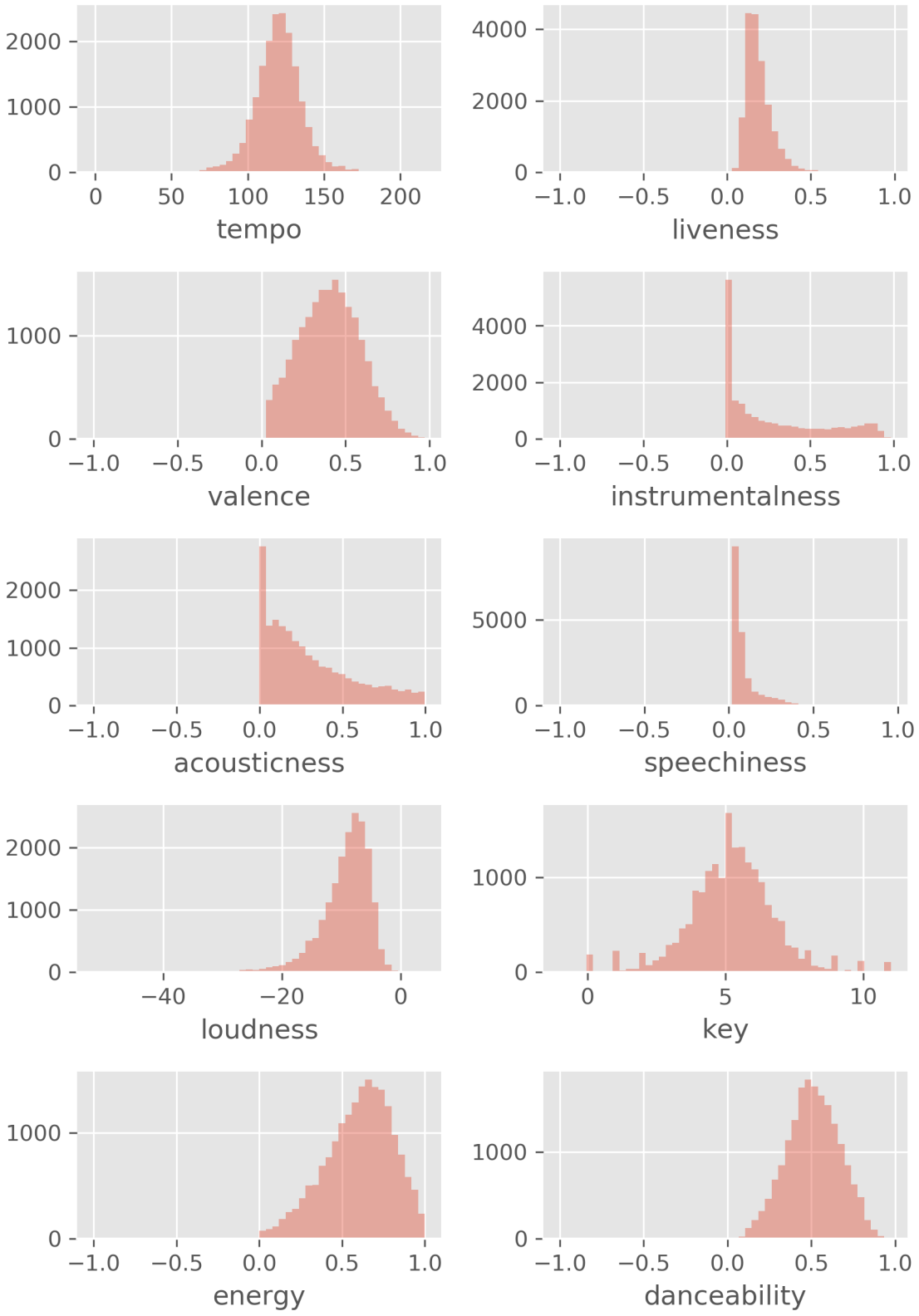

Figure 5: Audio feature distributions of the albums in the dataset.

ever, collaborative-based models can become problematic when rating data is not available for a song, and when there is not enough diversity or overlap in reviewer profiles because individual viewpoints can dominant the model. Finally, hybrid-based models for music recommendations have overcome some of these prior obstacles by unifying collaborative and content-based data, ultimately creating more accurate recommendations (**?**).

The P4KxSpotify dataset could serve as a valuable tool for training and evaluating music recommender systems, as it provides a strong set of item content features that can be utilized for all of the models previously mentioned with minimal adjustment.

For content-based models, the musical features of songs aggregated from Spotify (such as danceability, energy, or key) could improve the accuracy of these types of models. Additionally, while the set of reviewer profiles from Pitchfork is not large enough to train a standalone collaborative-based model, the addition of this information to a content-based model built around audio features could be a useful way of contextualizing the recommendations.

For metadata-based models, the information gleaned from the Pitchfork side of the dataset could be useful for making high-level recommendations to users (e.g., "You liked this album; so did reviewer Ian Cohen. Here a few other albums he really liked").

Finally, all of the dataset's features in combination with another source of large-scale reviewer profile information can be utilized to research more accurate hybrid-based models, such as probabilistic generative systems that have historically needed more sources of aggregated data for song ratings, content, and metadata.

Applications in Computational Musicology

Within the field of computational musicology, we see two clear applications in which our dataset might be useful.

First, one area of scholarship within computational musicology is concerned with understanding how music has changed over time. To accomplish this, researchers have examined how musical features have changed longitudinally, finding that generations tend to favor the genres of their respective youths (**?**) and that popular music changes in rapid bursts (**?**). Popular streaming services have also contributed to this, with projects such as Pandora's Music Genome Project[23] attempting to categorize music computationally. Our dataset contains 10 common musical features that are used to describe the music for albums that are considered to be culturally important in the United States, and spans a timeframe of over 60 years. Thus, our dataset might be useful to researchers interested in examining how popular music has changed from the mid-20th century to now.

Second, there is research in computational musicology focused on how to use musical features in conjunction with machine learning techniques such as neural networks to generate novel music using computers (e.g., (**?**; **?**)). Again, our database comprises over 18,000 records, making it a useful data source for research focused on this particular application of machine learning methodologies.

Conclusion

In this paper, we presented our novel dataset P4KxSpotify. Our dataset consists of the review data for music albums from Pitchfork and the musical attributes from the albums from Spotify's public API. We described the process for comprising this dataset from two sources (Pitchfork and Spotify), which might offer a replicable methodology for creating useful datasets in other contexts. Then, we presented basic descriptive statistics of our dataset. Finally, we offered several possible applications across different research contexts to highlight our dataset's potential usefulness to a variety of research endeavors, such as recommendation systems or in the field of computational musicology.

Acknowledgements

The authors wish to thank Tom Mullen of Washed Up Emo for inspiring this work with his Instagram post critiquing Pitchfork,[24] Kyle Gach for sharing Soloman Goldfarb's blog post with us, Brianna Dym for her feedback on an early draft, Casey Fiesler for her suggestions regarding the ethics of collecting data, and all of the musicians listened to while creating this dataset and writing this paper.

[23] https://www.pandora.com/about/mgp

[24] https://www.instagram.com/p/BMAbBslAHgx/