Measuring the accuracy of music genre classifier models using cross-collection evaluation

Borna Salarian[1]
Supervisors: Cynthia Liem[1] **Jaehun Kim**[1]
[1]EEMCS, Delft University of Technology, The Netherlands
b.salarian@student.tudelft.nl, , c.c.s.liem@tudelft.nl , J.H.Kim@tudelft.nl

Abstract

Working with trustworthy classifier models is important to the field of music information retrieval. However studies have shown some of the classifier models may not be as trustworthy as they appear. In this paper, we examine three of such classifiers available in the Essentia toolkit that have been evaluated using cross-validation, and measure the accuracy of these genre classifiers using cross-collection methods. We define a methodology inspired by other research in information retrieval to compare the output of the classifiers to an independent set of ground truth annotations that were the result of collaboration between the users of Last.fm. The classifiers were evaluated on 341 songs from the Muziekweb collection, and the results show that the classifiers performed worse than their cross-validation results.

1 Introduction

Music information retrieval (MIR) is a growing field of research that aims at developing computational tools and methods for analyzing and processing music [6]. Similar to all information retrieval tasks, MIR is a highly experimental discipline. Evaluation experiments on different MIR methods and careful interpretation of their results are essential to the field. Therefore it is important to work on and improve the methods used to evaluate different MIR methods [12].

It would be useful if researchers of the field could share all the music-related data with each other, leading to more trustworthy tools and methods as they can be replicated using the same data that they were trained on. However that is not the case since many large-scale music audio datasets can not be legally shared. There are only a few publicly available datasets that researchers in the field can all have access to and run experiments on. As a result, efforts have been made to locally pre-compute music audio descriptors and make them publicly available as part of research datasets [4,9]. These descriptors have been used to train machine learning pipelines such as the ones in Essentia[1] which will be further investigated in this paper.

[1]https://essentia.upf.edu/

The genre classifying machine learning pipelines available as part of Essentia's toolkit were shown to perform well using cross-validation techniques. However, when Bogdanov et all. in [3] tested some of these pipelines on an independent set of ground truths, the results were not as promising; with an accuracy between 43-58% achieved by some of the better performing pipelines.

The findings by the researchers [3,7,10] suggest that commonly used genre classifying machine learning pipelines are not as trustworthy as they appeared to be. In this paper, we aim to measure the accuracy of genre classifier models using cross-collection evaluation. How accurate are the genre classifying models included in Essentia when evaluated against an independent source of ground truth?

In the remainder of this paper, we further discuss Evaluation and its obstacles in section 2. Then in section 3 we will introduce our datasets and discuss our methodology. In section 4 we apply the methodology and the results are shown and discussed in section 5. In the penultimate section 6 we will discuss potential ethical implications. Lastly we will conclude the research and discuss some ideas for future work in section 7.

2 Background

In this section we motivate why evaluation is important and discuss some obstacles that are present in MIR evaluations.

2.1 Evaluation in MIR

Evaluation is one of the grand challenges not only in music information retrieval, but also more generally in the field of information retrieval. In September of 2002, a workshop made up of leading IR researchers recognized "Improved objective evaluation" as one of the seven major challenges in IR [1]. When a second workshop was held in 2012, the issue of evaluation was described as a "perennial" issue in IR and held its place as one of the major issues in IR [2].It is no surprise then that for music information retrieval (being a subbranch of the information retrieval field) evaluation is one of its major challenges.

Typical evaluation experiments in IR use a test collection used in conjunction with evaluation measures [8]. This follows the traditional Cranfiel paradigm based on experiments run by Cleverdon in the field of information retrieval in the

1960s [5]. A test collection has the three following components [13]:

- A collection of documents.
- A set of information needs. (also referred to as queries.)
- A set of relevance judgments. (also referred to as ground truth.)

Usually in an IR research, first a set of tasks would be identified. Then a document collection and a set of information needs are selected to simulate the potential requests of the user. When the system has evaluated the result of that query set and document collection (also known as a run), then these results are evaluated using several of the aforementioned evaluation measures [13].

2.2 The ground truth problem

An issue that is faced in MIR and more specifically when it comes to genre classification is the problem of a well defined ground truth. The definition of genres can be unclear, as it becomes quite subjective to determine which songs belong to which genre(s).For example a genre like Pop is very dynamic and ever changing with the changes of culture which makes it difficult to determine. We discuss how we approached this issue in the methodology section.

2.3 The collection problem

One of the more unique problems faced in MIR in particular, is the lack of a proper test collection to run evaluation experiments with. Music collections are not easily shareable due to copy right laws, and this leads to many issues. Music collections are often either too small or too biased [13] leading to flawed results on pipelines trained on them. Some researchers used their own personal music collections in order to train pipelines, which adds to the bias of the collection. In total there is a lack of a standard procedure when producing and experimenting on these collections. [13]

Efforts have been made by collections such as FMA [2] to make publicly available and royalty-free music collections. However in FMAs case the songs are often independent or lesser-known bands and therefore do not properly reflect the more general music scene and can lead to biases in tools that are trained with this collection.

3 Cross-collection evaluation methodology

The Cranfield paradigm as explained in section 2.1 is followed to perform evaluation experiments. The process is slightly modified however; instead of documents being retrieved in response to a query, systems provide tags for the query itself [13]. We will need three components to perform the evaluation. We define cross-collection evaluation to be an evaluation of a query (in this case genre classifiers) on a collection of documents (the music database) with an independent source of ground truth (music annotations generated by community.). The origin of the collection of documents is also different than the collection that the classifiers were trained on, making a further distinction with other evaluation methods such as holdout validation in which a part of the collection is used to verify the accuracy of the model.

3.1 Evaluation strategy

There are various evaluation strategies when it comes to comparing the estimation of a classifier and the ground truth annotations. One of the main challenges of cross-dataset analysis is a lack of a shared vocabulary [11]. Different music datasets and models do not necessarily share the same categories of music genre, making the direct mapping of classes between the classifiers and the ground truth not always an option.

The mapping of classes between the classifier and the ground truth can result in the following instances:

1. A class in the classifier can directly map to a class in the ground truth.
2. Multiple classes in the classifier can map to one class in the ground truth.
3. A class in the classifier can map to multiple classes in the ground truth.
4. A class in the classifier can not be mapped to any class in the ground truth.
5. A class in the ground truth can not be mapped to any class in the classifier.

For each case we need to map accordingly. In case 5 there is no mapping that can be done to the classifier while in case 4 there is a class that is known by the classifier but not known by the ground truth. In the cases that there are multiple classes mapping to one class we define a genre hierarchy that each of the sub-genres can be mapped to an overarching genre. We talk more about this genre mapping in section 4.

We will use all the recordings that are available in the ground truth set. If there are classes in the ground truth that do not match to a classifier, we will count that as a misclassification (e.g if a song is classical and the classifier does not have that as a label, we will count it as a misclassification.).

Songs in our ground truth set can be annotated with multiple genres (e.g. a song can be both pop and rock.). The genre classifiers however have only one genre estimation as an output. As a consequence, if the classifier correctly matches any of the ground truth annotations we count it as a correct classification.

3.2 Music Audio Dataset

A collection of independent data that differ from the data the classifiers were trained on is needed to perform cross-collection evaluation experiments as previously defined. For this research we partnered with Muziekweb[3] to procure an original corpus of data. Muziekweb is a music library based in Netherlands that has built a collection of music that has been released in the Netherlands since 1961.

To build our collection, a random number generator was used to randomly choose an album for each desired genre. This process was repeated till 50 songs were obtained for each of the following genres: Blues, Country, Classical, Disco,

[2] https://freemusicarchive.org/home

[3] https://www.muziekweb.nl/

Classifier	Accuracy	Size	Number of genres
ROS	87.56	400	8
GTZAN	75.53	1000	10
MABDS	60.25	1886	11

Table 1: Cross-validation accuracies (%) of classifier models

Dance, Electronic, Funk, Hip-Hop, Latin, Metal, Rock, Reggae, Pop, Jazz and an assortment of "World" music from Africa, India, Middle East, Asia, Western Europe, Southern and Southeast Europe summing up to a corpus of 1050 songs in total. (some genres including Classical, Funk, Dance, World, Electronic and Latin had 100 songs obtained each.)

3.3 Genre classifier models

Following along the Cranefield paradigm, we use a task to annotate the collection of songs. The task here is genre classification, and we examine the three following genre classifiers that are bundled with Essentia:

- **AcousticBrainz Rosamerica Collection (ROS)** A collection of 400 tracks for 8 genres (50 tracks per genre). The genres included are classical, dance, hip-hop, jazz, pop, rhythm blues, rock speech.
- **GTZAN Genre Collection (GTZAN)**[4] A collection of 1000 tracks for 10 music genres (100 per genre.). The genres included are blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock.
- **Music Audio Benchmark Data Set (MABDS)**[5] A collection of 1886 tracks made by TU Dortmund. The genres included are alternative, electronic, funk/soul/rnb, pop, rock, blues, folk/country, jazz, and rap/hiphop.

The cross-validation accuracy of the models as reported on AccousticBrainz can be seen under Table 1.

3.4 Independent ground truth

As part of the cross-collection evaluation, a set of independently sourced ground truth is needed. We use the tags generated by the community of Last.Fm [6]. The tags are voted by the users of the website and given a large enough number of votes to provide a good baseline as a source of ground truth. Not all tags from the website are accepted, as steps are taken to normalize the annotations received. We delve into more detail about the process in section 4.

3.5 Evaluation measures

As the last step of the evaluation methodology, the results gained from the query done on the collection are evaluated against the ground truth. If the genre classifier has correctly identified at least one tag that matches the ground truth set, it is counted as correct and adds to the accuracy percentage of the model. The accuracy of a model is defined as the percentage of correctly identified tags. We will use the f-score of the confusion matrix as a measure of this accuracy.

[4]http://marsyas.info/downloads/datasets.html

[5]https://www-ai.cs.tu-dortmund.de/audio.html

[6]https://www.last.fm/home

4 Evaluation of genre classifying models

In this section the methodology described before is applied. We use the Muziekweb data as our collection of documents, the genre classifier results as our query and the collected annotations from last.Fm are our independent ground truth.

4.1 Gathering ground truth

There are no restrictions on the genre tags that can be created by users on Last.fm. Though most tags correspond to a known genre, there are tags that do not correspond to any particular genre or provide any information. Keeping this in mind, each track from the collection was first matched to a song in the Last.fm database and then all the tags for that track were retrieved. The tags have a weight assigned to them, where the most commonly applied tags get a weight of 100. The tags that had a weight lower than 30 were discarded. Since there is no preset set of tags that users have to choose, this leads to many different naming conventions for genres and sub-genres (e.g.). In total there were 670 unique tags, and these need to be matched to the output of the classifier models.

To accomplish this we need to define a genre tree. This genre tree has "super-genres" that further part in sub-genres. We used the list of genres from Wikipedia [7] in combination with the genre tree from musicmap [8] to define our super-genres. The website attempts to form a genealogy of popular music genres, including their history and relations. Whenever a specific tag did not match any of the defined sub-genres on Wikipedia, musicmap was used to assign that tag to a known genre. Wikipedia categorizes popular music into 10 genres, plus 4 regional genres and avant-garde music. We discarded the art and religious categories, and combined all folk music into the genre "folk". Furthermore Caribbean and Latin songs were combined into one category, leaving us with 15 categories in total: African, Asian, Avant-garde, Blues, Caribbean and Latin, Classical, Country, Easy listening, Electronic, Folk, Hip-hop, Jazz, Pop, 'Rhythm Blues' and Rock.

Of the 1050 tracks that were in the original collection, only 385 of them were tagged by the users. This can be attributed to the fact that though the original dataset was fairly homogeneously distributed across genres, it does not represent the popularity of each genre. This resulted in a set of ground truth that was skewed towards more popular songs.

After each track was mapped to their corresponding genre, the 385 annotated tracks were reduced to 341. The full distribution of the percentage of songs across genres can be seen under table 2. Electronic and Rock music at 20.33% and 14.43% respectively had the largest share of the set, whereas Blues and Avant-garde songs had the lowest at 1.83% and 2.03% songs each. Most notably no Classical songs were matched with the songs available on Last.fm. This is mainly due to the fact that the titles of classical songs were mostly in Dutch and also did not follow the same naming conventions that was used on Last.fm (e.g. Symfonie nr.3 as opposed to symphony no.3). This lack of standard naming hindered the

[7]https://en.wikipedia.org/wiki/List$_of_music_genres_and_styles$

[8]https://musicmap.info/

Ground Truth	GTZAN	MABDS	ROS
african	-	-	-
asian	-	-	-
avant-garde	-	-	-
blues	blues	blues	-
caribbean	reggae	-	-
classical	classical	-	classical
country	country	folk/country	-
easy listening	-	-	-
electronic	disco	electronic	dance
folk	-	folk/country	-
hip hop	hip-hop	rap/hiphop	hip-hop
jazz	jazz	jazz	jazz
pop	pop	pop	pop
rhythm and blues	-	funk/soul/rnb	rnb
rock	rock, metal	rock, alternative	rock

Table 2: Mapping of ground truth genres to classifier classes.

process that was used to search the Last.fm database for a match.

Though the ground truth is skewed towards more popular songs, this also reflects the reality of what the models are expected to classify, as not all genres have the similar amount of following. A complete distribution can be seen as percentage of songs per genre in figure 1.

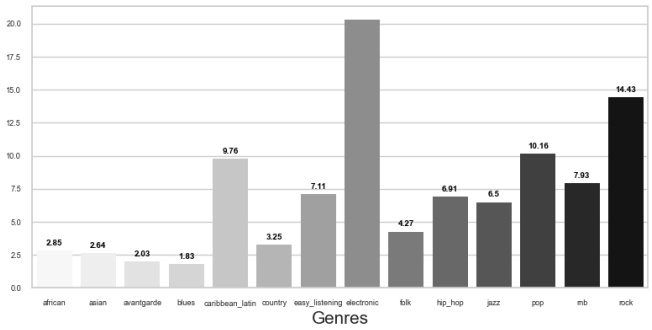

Figure 1: Percentage of songs per genre

4.2 Mapping genres

The classes of each genre classifier were mapped to the ground truth classes as can be seen under table 3.

As per cases discussed in our evaluation strategy, some classes such as the 'Folk/Country' class for MABDS matched to more than one class in the ground truth set (case 3), the 'rock' and 'metal' classes from GTZAN were mapped to one class (case 2), while some others did not have a match (case 4). These include the 'speech' classifier for ROS since speech is not a music genre, and 'international' and 'vocal' since they are not specific enough. we mapped 'dance' and 'disco' from ROS and GTZAN respectively to the 'electronic' class in ground truth as they were the closest match to it.

5 Results and discussion

The experiments were performed on 341 recordings. We measured the f-score of the each classifier. The f-score is a measure of accuracy in binary classification, each classifier either classified each song correctly or not. Furthermore there is also the weighted average f-score, which takes the imbalances of the dataset into account with regards to the number of labels per genre. The full result of the f-score of classifiers per genre are shown under table 4. The support column is the number of the correct samples available per genre.

The results are in direct conflict with the cross-validation accuracy as seen in table 1. None of the classifiers performed better than 40%, and GTZAN performed far worse than what cross-validation testing had suggested for its performance. In fact GTZAN classified all but 7 of the songs as Jazz, most of them with the same exact value of certainty. The rest of the 7 songs were classified as classical songs, which were absent in the ground truth set.

The MABDS classifier, although not as much as GTZAN, had a heavy bias towards a genre. It recognized 68% of the dataset as electronic. During the training of this classifier only 6% of the dataset used to train the model were electronic songs.

The ROS model did not have as an extreme bias towards a particular genre as other classifiers. It mainly confused electronic with hip-hop and rnb, mistaking 26% and 20% of the electronic songs for the aforementioned genres respectively. The full confusion matrix for each of the classifiers can be seen in figures 2-4.

In general the classifiers performed far worse than when evaluated using cross-validation methods.Consistent with the accuracy expected from table 1, ROS performed the best overall by the weighted average accuracy measure, which takes the imbalances of the dataset into account by assigning weights that depend on the number of true labels per genre. This result is interesting given that ROS had the lowest size of training data compared to the two other classifiers.

If only the unweighted f-score is given focus, the MABDS dataset performed slightly better than ROS, becoming the most accurate classifier. This is due to the fact that the ground truth dataset is not evenly spread across the genres, and the bias of MABDS towards electronic songs works in its favor since most of the songs in the ground truth set were electronic songs.

GTZAN performed by far the worst, with the weighted average f-score of only 0.01. It has to be noted if the distribution of the songs differed, the classifier might have performed better. Still in this sample of songs it failed to have any meaningful classification.

6 Responsible Research

The risk of classifiers incorrectly classifying information is a problem that in recent years is becoming more prevalent as machine learning techniques are applied to daily issues. These can lead to injustices that were paved by a lack of attention to the dataset that the machines were trained on [14]. In this paper the amount of data available does not represent

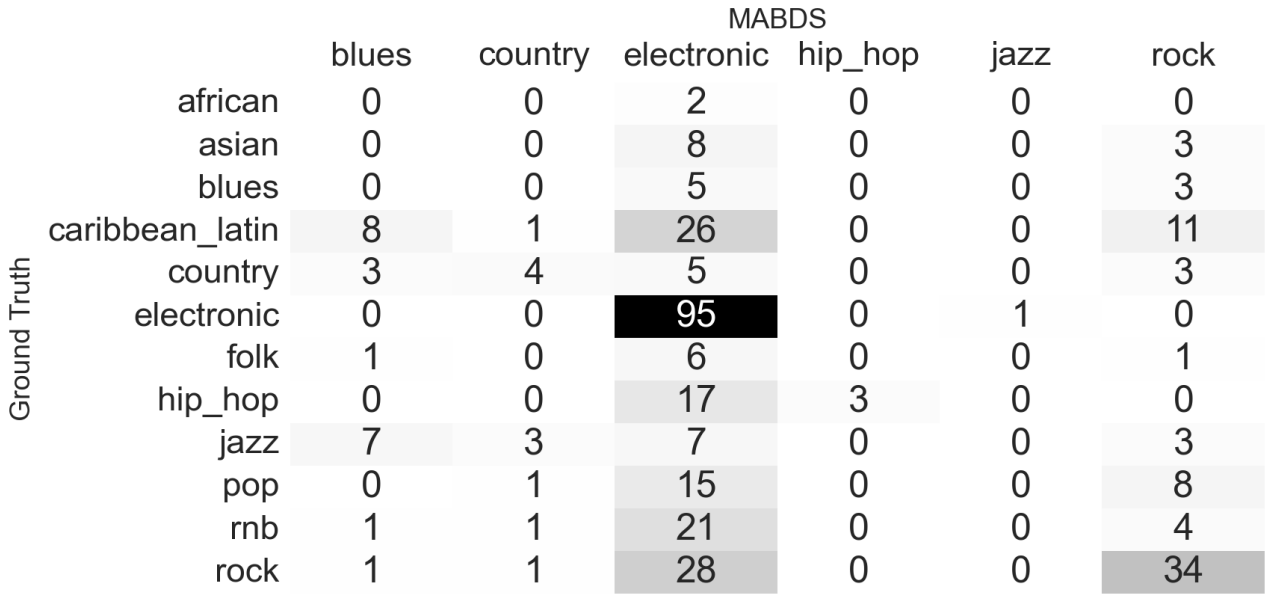

Figure 2: Confusion matrix of the MABDS classifier

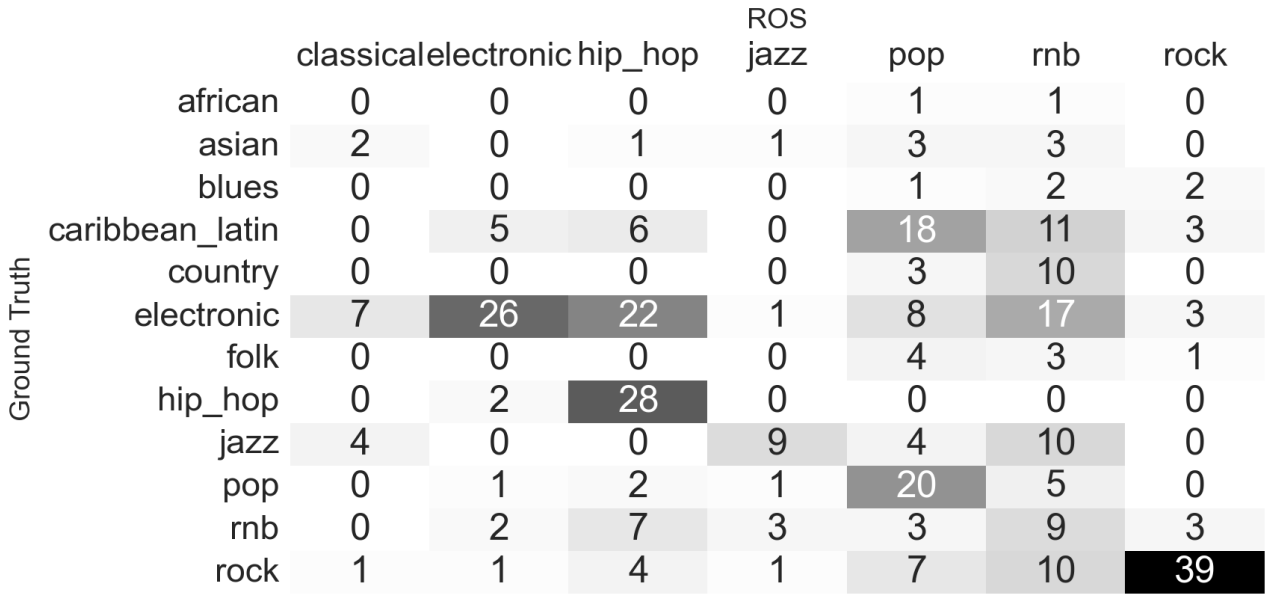

Figure 3: Confusion matrix of the ROS classifier

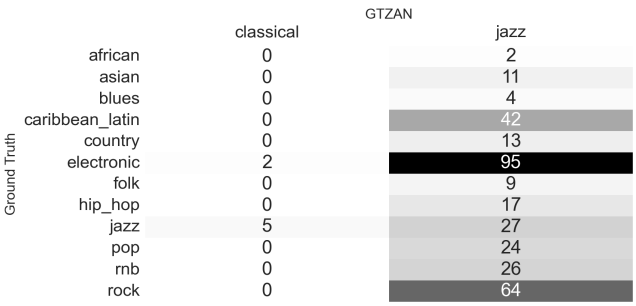

Figure 4: Confusion matrix of the GTZAN

Ground Truth	GTZAN	MABDS	ROS	support
african	0	0	0	2
asian	0	0	0	10
avant-garde	-	-	-	-
blues	0	0	0	5
caribbean	0	0	0	43
classical	-	-	-	-
country	0	0.31	0	13
easy listening	-	-	-	-
electronic	0	0.57	0.43	84
folk	0	0	0	8
hip hop	0	0.26	0.56	30
jazz	0.15	0	0.42	27
pop	0	0	0.40	29
rhythm and blues	0	0	0.17	27
rock	0	0.51	0.68	63
Accuracy	0.08	0.40	0.38	341
Weighted Avg.	0.01	0.29	0.36	341

Table 3: Accuracy (F1 score) of classifiers per genre.

the true spread of music across genres, and this can lead to issues if the methodology is applied on other classifiers without expanding the data first.

It also has to be noted that these classifiers mainly focus on popular songs, and so many songs from across the world are not properly classified with these tools.

7 Conclusions and Future Work

In this research we discussed the importance of trustworthy music feature extraction tools, and the obstacles that arise with evaluating the tools. We defined a methodology to apply information retrieval evaluation techniques on three of Essentia's genre classifiers, by collecting music and ground truth from independent sources of data and then comparing the results of the classifiers with the ground truth.

The methodology was applied on 341 songs, and the results showed the classifiers performing worse than when evaluated using cross-validation techniques. The better performing models had an accuracy of 29-36%, while the worst one performed only at 1% accuracy.

These findings show that cross-validation techniques may not be enough to show the true performance of genre classifiers, and thus other methods of evaluation are necessary if one were to find the true accuracy of classifier models. More research is needed to design classifiers that more meaningfully extract musical information and so are in turn more robust.

In the future, the proposed methodology and tools generated by this paper can be easily expanded upon, both in terms of evaluating the same genres with much more data and also evaluating other classifiers. Essentia has many more classifiers that can be evaluated using the same methodology discussed here albeit with a few tweaks. Given the constrains of this research, the validity of the results suffer from the rather small sample size. If one were to repeat the methodology here with a much larger amount of data more insightful results can be acquired.

References

[1] James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J Harper, et al. Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst, september 2002. In *ACM SIGIR Forum*, volume 37, pages 31–47. ACM New York, NY, USA, 2003.

[2] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. In *ACM SIGIR Forum*, volume 46, pages 2–32. ACM New York, NY, USA, 2012.

[3] Dmitry Bogdanov, Alastair Porter, Herrera Boyer, Xavier Serra, et al. Cross-collection evaluation for music classification tasks. In *Devaney J, Mandel MI, Turnbull D, Tzanetakis G, editors. ISMIR 2016. Proceedings*

of the 17th International Society for Music Information Retrieval Conference; 2016 Aug 7-11; New York City (NY).[Canada]: ISMIR; 2016. p. 379-85. International Society for Music Information Retrieval (ISMIR), 2016.

[4] Dmitry Bogdanov, Alastair Porter, Julián Urbano, and Hendrik Schreiber. The mediaeval 2017 acousticbrainz genre task: Content-based music genre recognition from multiple sources. CEUR Workshop Proceedings, 2017.

[5] Cyril W Cleverdon. The significance of the cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 1991.

[6] Joe Futrelle and J Stephen Downie. Interdisciplinary communities and research issues in music information retrieval. In *ISMIR*, volume 2, pages 215–221, 2002.

[7] Cynthia CS Liem and Chris Mostert. Can't trust the feeling? how open data reveals unexpected behavior of high-level music descriptors. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020)*, 2020.

[8] Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010.

[9] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *ISMIR*, pages 469–474. Citeseer, 2012.

[10] Bob L Sturm. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

[11] Tatiana Tommasi and Tinne Tuytelaars. A testbed for cross-dataset analysis. In *European Conference on Computer Vision*, pages 18–31. Springer, 2014.

[12] Julián Urbano, Dmitry Bogdanov, Herrera Boyer, Emilia Gómez Gutiérrez, Xavier Serra, et al. What is the effect of audio quality on the robustness of mfccs and chroma features? In *Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2014); 2014 Oct 27-31; Taipei, Taiwan.[place unknown]: International Society for Music Information Retrieval; 2014. p. 573-578.* International Society for Music Information Retrieval (ISMIR), 2014.

[13] Julián Urbano, Markus Schedl, and Xavier Serra. Evaluation in music information retrieval. *Journal of Intelligent Information Systems*, 41(3):345–369, 2013.

[14] A Zimmerman, Elena Di Rosa, and Hochan Kim. Technology can't fix algorithmic injustice. *Boston Review*, 2020.