TAG ME UP LAST.FM

Multi-class lyrics classification, a Large Language Model approach

Sangeeths Chandrakumar
Fachhochschule Graubünden Switzerland
sangeeths.chandrakumar@stud.fhgr.ch
ORCID 0009-0007-0403-4661

Florian Klessascheck
Fachhochschule Graubünden Switzerland
florian.klessascheck@stud.fhgr.ch
ORCID 0009-0000-0709-7974

Adrian Joost
Fachhochschule Graubünden Switzerland
adrian.joost@stud.fhgr.ch
ORCID 0009-0002-0950-0119

Ana Petrus
Fachhochschule Graubünden / scivia GmbH Switzerland
ana.petrus@fhgr.ch
ORCID 0000-0002-0928-8894

***Abstract* – The automatic classification of music tracks according to their lyrics represents an innovative approach for music streaming services. Tagging systems assist users in discovering new music that aligns with their preferences, thereby enhancing satisfaction with the service. Previous research has employed residual neural networks for the classification of music based on their spectrogram. This paper investigates the potential of classifying tracks based solely on their lyrics. To this end, 48'000 tracks with their lyrics and tags were extracted from genius.com and Last.fm. From this dataset, 429 distinct tags were identified for evaluation purposes. Using this dataset, a Mistral-7b-instruct-v2 model was trained and evaluated, demonstrating classification scores of up to 80%. The results indicate that lyrics can serve as a reliable indicator for certain tags.**

***Keywords* – Large Language Models, classification, tagging, music, cataloguing**

I. INTRODUCTION

Last.fm is a website that enables users to add tags via a process known as collaborative tagging [1]. Those tags include, but are not limited to, genre, language and the overall feeling (or “vibe”) of the song. For the purposes of this research, the lyrics and their respective tags of 48'000 songs from popular (2024) artists were extracted from genius.com and last.fm.

II. STATE OF RESEARCH

The automatic tagging of audio sequences based on machine learning is a well-researched task [2]. For example, in their work "Audio tagging with noisy labels and minimal supervision" [3] Fonseca et al. focus on classifying urban sounds sources and music genres, using spectrum analysis in combination with recurrent neural networks.

Newer research suggests [4] that using lyrics and natural language processing yields better results on genre and "vibe" tagging of music.

III. RESEARCH QUESTION AND METHODOLOGY

The aforementioned works do not employ the use of large language models and are limited to a single prediction for the classification of a single lyric text. This paper builds upon the concept of tagging with the aid of large language models. Consequently, it seeks to answer the following question:

How well can a fine-tuned large language model perform a multi-class audio tag classification based on lyrics and the primary artist’s name alone?

The primary artist’s name was included to reflect a real-world scenario in which the artist responsible for a particular song is known. From the dataset that we have compiled, we have manually selected 429 of the approximately 4’000 tags that are present. The focus of this research is exclusively on the aforementioned tags. Any other tags will be disregarded, even if they are

SRDD 2024: Swiss Research Data Day, Chur (Switzerland).

DOI: 10.nnnn/nnnnnnn.nnnnnnn

present in the predictions. For the purposes of curation, tags that had been observed less than three times across the entire dataset, or that were of low quality (for example, labelled as “good” or “trash”), were filtered out. Any sub-genres or duplicate tags present in a single observation were merged into a single tag.

With regard to the prediction of tags, the mistral-7b-instruct-v2 model was selected, on the grounds of its excellent instruction cohesion and the availability of information on the fine tuning [5] of mistral models.

The initial dataset was divided into three subsets: a training set comprising 66% of the data, a test set comprising 17% of the data, and an evaluation set comprising 17% of the data. The fine-tuning was conducted in accordance with a guide [5] utilising Parameter-Efficient Fine-Tuning (PEFT) [6] with Low-Rank Adaptation (LoRA) [7], with the objective of reducing the computational resource requirements. Furthermore, in order to limit resource usage, all lyrics in the train and test splits exceeding 2’000 tokens were excluded from the data set, resulting in the exclusion of 34 songs (approximately 0.085%) in total.

To assess the impact of fine-tuning, two baselines were established using the unmodified base model. One baseline was provided with all curated tags, while the other was conducted blind, where the model was only presented with the identical prompt as the fine-tuned model:

```
"[INST]«SYS»Tag the song based on the
lyrics, only respond in json {"tags":
[]}«/SYS»\n[<artist name>]\n<lyrics>
[/INST]"
```

Due to a lack of resources, two evaluations were conducted. The first evaluation used a fine-tuned model and included 24’000 songs. The second evaluation used the previously described 8’000 songs and compared the fine-tuned model to the base model. However, due to an oversight in data preparation, songs without any tags after filtering were excluded from the evaluation set. This resulted in a reduction of 1’945 songs, leaving 22’055 songs in the first evaluation set.

Confusion scores were calculated on a per-tag basis. For each song, all curated tags were checked and scored in accordance with the following scheme.

A. *Evaluation Strategy*

This paper asserts the quality of the model by its precision, because the focus of classification is that a predicted tag is correct, not that the model finds all tags.

As the TN (true negative) count is often significantly higher in multi-class classification, the evaluation focused instead on precision, recall and F1 score.

Table 1: Confusion Scoring Strategy

	Tag in label	Tag in prediction
True positive	True	True
False positive	False	True
False negative	True	False
True negative	False	False

IV. Results

Accuracy and specificity were included but as shown in Appendix Figure 1, were heavily influenced by the high TN count. Based on this, looking at the tags ranked by precision, Appendix Figure 2 suggests that the genres country, rap and hip-hop can be well identified by their lyrics. Tags such as electronic, rhythm and blues or trap exhibited lower predictive accuracy.

For the overall best tags (Appendix Figure 3), the model performs well in predicting the languages (Appendix Figure 4) of the songs. The more important precision score is approximately 70% for the top 20 tags in general and by language. Some false positive and false negative songs[1] can be attributed to incorrect or incomplete last.fm tags. In these cases, the model correctly predicted the tag, but it was not included in the labels. For example, there are songs[2] without German lyrics[3] that are wrongly tagged as "German".

A. *Model Performance*

Figure 1 illustrates that the fine-tuned model exhibits the most optimal performance overall. The elevated precision scores can be attributed to the exclusion of any non-curated tags, even when accounting for false positives. Notably, the baseline that was provided with all available tags demonstrated a less favourable outcome than the baseline that was not assisted.

Run	Name	F1	Recall	Precision
finetuned	Top 20	63.14%	65.28%	61.12%
	Count ≥ 500	64.84%	67.98%	61.97%
	Count ≥ 250	63.83%	67.02%	60.93%
	Total	50.95%	46.14%	56.89%
baseline-tags	Top 20	38.67%	28.83%	58.72%
	Count ≥ 500	41.30%	31.42%	60.23%
	Count ≥ 250	39.61%	29.81%	59.02%
	Total	26.75%	21.73%	34.79%
baseline-blind	Top 20	41.91%	30.89%	65.13%
	Count ≥ 500	46.59%	35.56%	67.52%
	Count ≥ 250	42.60%	31.54%	65.60%
	Total	27.99%	19.81%	47.68%

Figure 1: Comparison of approaches

[1] out of scope of this paper, see https://github.com/Aeolin/tag-me-up-last-fm/blob/master/evaluation_results/confusion/cm_record_including_songs.json for further investigation

[2] https://www.last.fm/music/Michael+Jackson/_/Happy+Birthday+Lisa

[3] https://genius.com/Michael-jackson-happy-birthday-lisa-lyrics

V. Discussion

Analysis of the results shows that certain tags, such as "rap", are very well suited for evaluation by large language models, while others, such as "electronic", are not. The reason may be that some, but not all, tags have a correlation with the lyrics. Considering this, we suggest that a combination of different predictors or approaches might yield even better results for tagging songs, for example natural language processing for lyrics-heavy tags and sound waves for melody-focused tags.

The baseline model with all available tags performed worse than the blind baseline model, which may stem from an overfilled prompt containing a list of all 429 tags.

During the evaluation, after training, we discovered incomplete and incorrect last.fm tags. It is therefore recommended that, when recreating the approach described in this paper, steps are taken to improve the quality of the dataset by filtering out any unwanted tags or songs with poor label quality.

However, the desired tags must be defined in accordance with the requirements. Given the large number of songs by the same artist in the dataset, it would be beneficial to use a more diverse dataset in terms of artist count, or to use only lyrics, in order to prevent overfitting due to the recognition of the artist's name.

REFERENCES

[1] Y.-X. Chen, S. Boring, and A. Butz, 'How last. fm illustrates the musical world: user behavior and relevant user-generated content.', in *Proceedings of the international workshop on visual interfaces to the social and semantic web*, Hong Kong, China, 2010. [Online]. Available: https://www.medien.ifi.lmu.de/pubdb/publications/pub/chen2010VISSW2/chen2010VISSW2.pdf

[2] J. Breebaart and M. F. McKinney, 'Features for Audio Classification', in *Algorithms in Ambient Intelligence*, vol. 2, W. F. J. Verhaegh, E. Aarts, and J. Korst, Eds., in Philips Research, vol. 2, Dordrecht: Springer Netherlands, 2004, pp. 113–129. doi: 10.1007/978-94-017-0703-9_6.

[3] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, and X. Serra, 'Audio tagging with noisy labels and minimal supervision', 2019, *arXiv*. doi: 10.48550/ARXIV.1906.02975.

[4] A. Tsaptsinos, 'Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network', 2017, *arXiv*. doi: 10.48550/ARXIV.1707.04678.

[5] H. Carroll, I. Dhanani, N. Khalil, V. Parashar, T. Fong, and Anarkoic, 'notebooks/mistral-finetune-own-data.ipynb at main · brevdev/notebooks', Features for Audio and Music Classification. Accessed: Apr. 24, 2024. [Online]. Available: https://github.com/brevdev/notebooks/blob/main/mistral-finetune-own-data.ipynb

[6] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, 'Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment', 2023, *arXiv*. doi: 10.48550/ARXIV.2312.12148.

[7] Y. Yu *et al.*, 'Low-rank Adaptation of Large Language Model Rescoring for Parameter-Efficient Speech Recognition', in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2023, pp. 1–8. doi: 10.1109/ASRU57964.2023.10389632.

ACKNOWLEDGEMENTS

This study was conducted as part of the course "CDS1091 Natural Language Processing", taught by Prof. Corsin Capol, in the context of BSc Computational and Data Science at the University of Applied Sciences of the Grisons (Fachhochschule Graubünden), Switzerland.

CONTRIBUTIONS

Sangeeths Chandrakumar: conceptualisation, data curation, formal analysis, investigation, methodology, software, validation, visualisation, writing – original draft

Florian Klessascheck: conceptualisation, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft

Adrian Joost: conceptualisation, writing – original draft

Ana Petrus: supervision, writing – review & editing

Appendix

Appendix Figures 1 through 4.

Tag / Predicted Count / Actual Count	Accuracy	Specificity	Precision	Recall	F1
hip-hop 7415 5770	84.6%	88.6%	73.9%	91.0%	68.9%
rap 7297 5626	85.3%	89.1%	74.3%	92.4%	70.1%
pop 6777 5326	79.5%	86.1%	63.7%	76.9%	53.5%
rock 4082 2974	83.4%	90.6%	54.1%	71.0%	44.3%
alternative 2562 1938	83.1%	92.1%	33.4%	41.4%	22.7%
indie 1461 1908	85.7%	96.1%	42.0%	30.2%	21.3%
trap 2595 1783	86.5%	93.5%	47.0%	65.2%	37.6%
dance 326 1627	88.2%	99.5%	64.8%	12.7%	11.8%
soul 312 1507	88.8%	99.5%	67.1%	12.7%	12.0%
electronic 1496 1489	88.7%	96.6%	51.8%	50.2%	34.2%
rhythm and blues 1798 1397	89.0%	96.0%	51.3%	62.8%	39.4%
alternative rock 893 1127	91.1%	98.5%	65.0%	51.0%	40.0%
indie pop 1290 1116	90.2%	97.2%	51.5%	55.6%	36.5%
metal 1229 971	91.5%	97.6%	57.7%	71.3%	46.8%
80s 803 954	91.5%	98.6%	61.3%	49.6%	37.7%
folk 1494 817	91.0%	96.6%	46.9%	78.6%	41.6%
country 652 692	93.9%	99.6%	87.9%	79.5%	71.6%
jazz 367 662	92.6%	99.6%	72.9%	33.4%	29.7%
hard rock 281 648	92.8%	99.8%	83.9%	35.5%	33.2%
christian 232 627	92.5%	99.7%	69.3%	25.2%	22.7%

Appendix Figure 1: Scores for top 20 tags by actual count, ordered by the actual count

Tag / Predicted Count / Actual Count	Precision	Recall	F1
country 652 692	87.9%	79.5%	71.6%
hard rock 281 648	83.9%	35.5%	33.2%
rap 7297 5626	74.3%	92.4%	70.0%
hip-hop 7415 5770	73.9%	91.0%	68.9%
jazz 367 662	72.9%	33.4%	29.7%
christian 232 627	69.3%	25.2%	22.7%
soul 312 1507	67.1%	12.7%	12.0%
alternative rock 893 1127	65.0%	51.0%	40.0%
dance 326 1627	64.8%	12.7%	11.8%
pop 6777 5326	63.7%	76.9%	53.5%
80s 803 954	61.3%	49.6%	37.7%
metal 1229 971	57.7%	71.3%	46.8%
rock 4082 2974	54.1%	71.0%	44.3%
electronic 1496 1489	51.8%	50.2%	34.2%
indie pop 1290 1116	51.5%	55.6%	36.5%
rhythm and blues 1798 1397	51.3%	62.8%	39.4%
trap 2595 1783	47.0%	65.2%	37.6%
folk 1494 817	46.8%	78.6%	41.5%
indie 1461 1908	42.0%	30.2%	21.3%
alternative 2562 1938	33.4%	41.4%	22.7%

Appendix Figure 2: Scores for top 20 tags by actual count, ordered by precision

Tag / Predicted Count / Actual Count			F1	Precision	Recall
romanian	190	194	93.4%	97.9%	95.4%
grunge	433	409	90.7%	93.0%	97.3%
grindcore	245	238	88.4%	94.5%	93.3%
industrial	304	309	74.3%	88.2%	82.5%
country	652	692	71.6%	87.9%	79.5%
k-pop	134	130	71.1%	83.1%	83.1%
rap	7297	5626	70.0%	74.3%	92.4%
hip-hop	7415	5770	68.9%	73.9%	91.0%
heavy metal	483	570	63.6%	85.8%	71.1%
nu metal	278	289	63.5%	79.6%	75.8%
christian rock	212	152	62.5%	67.8%	88.8%
60s	148	152	57.4%	77.2%	69.1%
pop	6777	5326	53.5%	63.7%	76.9%
hyphy	125	167	53.3%	85.2%	58.7%
thrash metal	372	539	51.2%	84.7%	56.4%
metal	1229	971	46.8%	57.7%	71.3%
rock	4082	2974	44.3%	54.1%	71.0%
j-pop	174	125	43.3%	52.0%	72.0%
gospel	218	450	43.2%	93.1%	44.7%
punk rock	315	382	41.9%	66.6%	53.1%

Appendix Figure 3: Scores for top 20 tags by F1 and > 100 actual counts, ordered by F1

Tag / Predicted Count / Actual Count			F1	Precision	Recall
romanian	190	194	93.4%	97.9%	95.4%
arabic	47	46	87.8%	93.5%	93.5%
k-pop	134	130	71.1%	83.1%	83.1%
norwegian	61	63	64.0%	80.0%	76.2%
turkish	66	56	53.2%	66.1%	73.2%
metal	1229	971	46.8%	57.7%	71.3%
korean	77	83	44.5%	64.5%	59.0%
indonesian	118	60	43.6%	47.2%	85.0%
j-pop	174	125	43.3%	52.0%	72.0%
spanish	294	246	37.3%	51.3%	57.7%
swedish	258	434	37.0%	75.0%	42.2%
azerbaijan	43	122	35.2%	100.0%	35.2%
russia	158	252	33.6%	65.2%	40.9%
italian	171	189	33.1%	54.0%	46.0%
french rap	354	511	29.9%	57.1%	38.6%
serbian	31	90	28.7%	87.1%	30.0%
german rap	62	143	26.7%	70.5%	30.1%
british	282	570	25.1%	63.2%	29.5%
danish	27	112	24.1%	100.0%	24.1%
japanese	165	211	22.5%	41.8%	32.7%

Appendix Figure 4: Scores for top 20 language related tags by F1