Voices of Dissent: A Multimodal Analysis of Protest Songs through Lyrics and Audio

Utsav Shekhar
IIIT Hyderabad
utsav.shekhar@research.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

Music has long served as a vehicle for political expression, with protest songs playing a central role in articulating dissent and mobilizing collective action. Yet, despite their cultural significance, the linguistic and acoustic signatures that define protest music remain understudied. We present a multimodal computational analysis of protest and non-protest songs spanning multiple decades. Using NLP and audio analysis, we identify the linguistic and musical features that differentiate protest songs. Instead of focusing on classification performance, we treat classification as a diagnostic tool to investigate these features and reveal broader patterns. **Protest songs are not just politically charged; they are acoustically and linguistically distinct, and we quantify how.**

1 Introduction

Protest songs have historically functioned as powerful tools for voicing dissent, mobilizing communities, and challenging dominant narratives. From anthems echoing through mass gatherings to quiet songs of resistance passed down across generations, protest music has consistently voiced the collective conscience. As demonstrated during Kenya's 2024 Gen Z-led protests, music holds a dualistic power serving both as a cultural artefact and a potent political tool for resistance and unity (Kirui, 2025).

Protest songs often transform personal struggles into shared narratives. During the U.S. Civil Rights Movement, We Shall Overcome became a symbol of unity and resilience (Conklin, 2014). In South Africa, anti-apartheid songs voiced resistance against systemic oppression (Drewett, 2003). India's anti-colonial movement used music to instill courage and national identity (Raha, 2018), while anti-war songs during the Vietnam era amplified global dissent. More recently, Turkey's Gezi Park protests (Bianchi, 2018) and Burkina Faso's pop-driven civic critique (Ouedraogo, 2018) illustrate the enduring mobilizing power of music in diverse political contexts.

While prior work has emphasized the cultural and social impact of protest music, the linguistic and acoustic features that distinguish protest songs from non-protest ones remain largely underexplored. Most existing studies focus on symbolic, thematic, or historical dimensions, with limited use of computational methods. One exception is (Miller, 1997), who manually annotated protest songs from 1963 to 1970 to analyze thematic patterns and stylistic features. However, such manual analyses limited in scope and scale fall short of capturing the full range of linguistic and acoustic markers that define protest music.

To address this gap, we present a multimodal computational analysis of protest music. We compile a dataset of protest songs from (Jiang and Jin, 2022), sourced via Wikidata, and pair it with a matched set of non-protest songs selected using GPT-4 inference (OpenAI, 2023), aligned by time period and ensuring genre diversity. Identifying what differentiates protest music from other forms illuminates how dissent is encoded in both language and sound, with implications for musicology, political communication, and digital activism.

2 Our Contributions

This work presents a comprehensive computational study of protest music through the following contributions:

- **A multimodal protest music dataset.** We curate a novel dataset of 446 protest and 370 non-protest songs spanning diverse genres, languages and decades. Each song includes full lyrics, 30 second audio excerpts, and source separated vocal/accompaniment tracks. Protest songs are sourced from Wikidata (Jiang and Jin, 2022), while non-protest songs are filtered via GPT inference (OpenAI, 2023).

- **Text-based classification.** We use multiple transformer-based embeddings for protest song classification, including both music informed and general purpose text architectures. Our comparative analysis shows that protest lyrics exhibit systematic and classifiable differences from non-protest songs.

- **Interpretable linguistic feature analysis.** We extract and analyze a diverse set of interpretable linguistic features to isolate the dimensions that distinguish protest lyrics from non-protest ones. Protest songs exhibit significantly higher repetition, lexical diversity, and sentiment polarity, among other stylistic differences.

- **Audio-based classification.** We evaluate a range of pretrained audio models both general-purpose and music specific for protest classification directly from raw audio. Vocal segments consistently yield higher performance than instrumental ones, underscoring the centrality of vocal expression in protest music.

- **Audio feature analysis.** We extract and analyze a range of interpretable audio features to investigate the auditory dimensions that distinguish protest songs from non-protest songs. Key features such as repetition, spectral rolloff, energy fluctuations etc extracted from librosa (McFee et al., 2015) library are used for comparative analysis. Also, we human-annotated perceptual audio features and found protest songs to be generally faster, more energetic, and less acoustic than non-protest songs

Source Separation. We decompose audio tracks into vocals and accompaniment to analyze whether protest signals are more strongly embedded in the lyrics or the musical arrangement. Each stem is classified independently to assess its contribution to protest prediction. Additionally, we conduct a controlled mixing experiment, combining protest vocals with non-protest accompaniment and vice versa, to quantify the influence of vocal and instrumental components on protest music classification.

3 Dataset

Our dataset consists of two primary categories: protest songs and non-protest songs. The protest songs were sourced from a list curated by (Jiang and Jin, 2022), which was itself compiled from Wikipedia and includes 459 tracks linked to various protest movements across different decades and regions. For each song in this collection, we obtained relevant metadata, Spotify and Wikipedia links, and retrieved lyrics using the Genius API[1]. Of these, lyrics were successfully extracted for 458 tracks, with only one track missing due to unavailability.

To construct a suitable non-protest comparison set, we curated a collection of 400 songs spanning a wide range of musical genres from roughly the same time periods as the protest songs. GPT-4 (OpenAI, 2023) inference was employed to ensure that these tracks were not associated with any social or political movements. Specifically, we used GPT's search functionality to identify popular songs from diverse genres, carefully maintaining a balanced distribution across both decades and musical styles. It was then manually verified that the songs are well spread across time and are not related to any protest. Through the same lyrics extraction pipeline used for protest songs, we successfully retrieved lyrics for 370 of the non-protest tracks.

The genre distribution across the two categories reveals some notable contrasts. In the protest set, pop (21.69%), rock (18.03%), and disco (16.06%) were the most prominent genres, followed by hip hop (14.93%), country (9.58%), reggae (9.30%), blues (5.35%), classical (2.54%), metal (2.25%), and jazz (0.28%). In contrast, the non-protest set was dominated by rock (27.91%) and metal (17.79%), with country (12.88%), hip hop (11.04%), pop (10.12%), reggae (7.36%), disco (5.21%), blues (3.07%), classical (3.07%), and jazz (1.53%) following behind. Genre labels for each song were derived using a music classification model fine-tuned on the GTZAN dataset.[2]

[1] (https://genius.com)

[2] https://huggingface.co/hungphan111/music_genres_classification-finetuned-gtzan-finetuned-gtzan

Audio availability posed certain limitations. For protest songs, we were able to locate publicly accessible audio for 330 of the 459 tracks, primarily through Spotify links. In the case of non-protest songs, audio was available for 355 tracks. These were retrieved using the Pytube library, which enabled us to extract audio from publicly available YouTube uploads. To ensure consistency in analysis, we used 30-second excerpts from each song. Since the beginning of many YouTube videos contains silence or low-volume intros, we extracted segments from the 15 to 45-second mark to capture audio-rich sections for more accurate processing. We acknowledge that the choice of non-protest songs can influence classification difficulty. Future work could construct more adversarial baselines (e.g., thematically similar but apolitical songs) to further probe the boundary between protest and non-protest music

Song Type	Initial Count	Lyrics	Audio
Protest	459	458	330
Non-Protest	400	370	355

Table 1: Dataset Summary

4 Methodology

4.1 Overview

We adopt a multimodal approach to characterize and classify protest music using both textual and audio representations. Our pipeline involves (1)Using only the textual part of the song (Lyrics) for analysis. (2) Using the audio part of the song for analysis (both vocals and accompaniment) (3) We also perform source separation to isolate vocals and accompaniment for analysis and 4) conduct human annotation to validate high-level musical differences. The annotated features such as repetition, ornamentation and melodic disjunctness were selected based on prior qualitative analysis by (Miller, 1997).

4.2 Linguistic Analysis

Our goal in this section is to investigate whether protest intent is reflected in the stylistic and structural properties of lyrics. To this end, we employ both deep contextual embeddings and interpretable linguistic features to identify the textual markers that differentiate protest songs from non-protest ones.

Embeddings. We encode each song's lyrics using several pretrained transformer models, including RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), DistilBERT (Sanh et al., 2020), and Veucci's Bert based lyrics-to-genre model[3]. RoBERTa, XLM-RoBERTa, and DistilBERT are language-driven models trained on general textual corpora, capturing syntactic and semantic properties. In contrast, Veucci's model is fine-tuned on genre-labeled lyrics and is more sensitive to musicality-related patterns. These models convert lyrics into fixed-size embeddings via mean pooling over the final-layer token representations. To accommodate lyrics exceeding the models' 512-token context window, we apply a sliding window approach with 50% overlap. Embeddings from each chunk are averaged to produce a single vector per song.

Rather than fine-tuning transformer models which risks overfitting on our limited dataset we use frozen embeddings as input features. These are evaluated using a range of classifiers: (1) Statistical models such as Logistic regression (Cox, 1958) , support vector machines (SVM) (Cortes and Vapnik, 1995), random forests for interpretability, and (2) lightweight neural models with trainable final layers, including a linear layer and a shallow multilayer perceptron (MLP) have been used. This setup enables a balanced comparison of language- and audio-based features across model complexity and generalization.

We employed an 80:20 train-test split to evaluate model performance. Additionally, we used k-fold cross-validation on the training set to enhance the robustness of our results and mitigate variance due to data partitioning. The final performance metrics reported are averaged F1 (Van Rijsbergen, 1979) metric scores computed across the folds, providing a more reliable estimate of the model's generalization capability.

[3] https://huggingface.co/Veucci/lyric-to-genre

Linguistic Features. In addition to deep embeddings, we extract a set of interpretable linguistic features designed to capture stylistic and structural properties of the lyrics. These include sentiment score, average line length, rhyme density, lexical density, the number of figurative expressions (such as metaphors and similes), unique word ratio, and repetition metrics such as unigram and bigram repetition. All features are normalized and used to train traditional classifiers, including logistic regression and ensemble-based models.

4.3 Audio Analysis

Deep Audio Representations. We extract fixed-size embeddings using pretrained audio models Contrastive Language-Audio Pretraining (CLAP) by (Elizalde et al., 2022), Hidden Unit BERT (HuBERT) by (Hsu et al., 2021), and Wave2Vec by (Baevski et al., 2020) without fine-tuning. CLAP captures joint language-musical cues, HuBERT focuses on speech-related features, and Wave2Vec, trained on raw audio, provides deeper speech representations. These embeddings serve as inputs to classifiers such as Support Vector Machines (SVM), Random Forest, and Multilayer Perceptrons (MLP), allowing for effective comparison between musicality and speech-driven representations. To ensure a fair and consistent evaluation, we adopt an 80:20 train-test split, stratified to maintain class balance across both sets. Within the training set, we perform k-fold cross-validation to account for variance in model performance due to data partitioning. Final results are reported as the average F1 score across folds on the held-out test set, providing a robust measure of classification effectiveness.

Audio Feature Extraction. We extract low-level audio features using Librosa (McFee et al., 2015) spectral flux, shimmer, and MFCCs which capture fine-grained aspects of timbre, dynamics, and texture. These audio features are used to train a logistic regression classifier, following the same setup as for linguistic features.

Human Annotation. To complement our computational analysis, we conducted human annotation on a subset of protest and non-protest songs (20 songs from each set were chosen for annotation) . Annotators rated musical attributes such as repetition, ornamentation, vocal roughness, melodic contour, and emotional delivery. These attributes were selected based on a qualitative framework from (Miller, 1997). The annotations were used to validate the directionality and salience of observed differences between the two categories. About 50 annotators participated in the experiment. Annotators were mostly from 20-25 age group and were students with mostly no formal musical training.

Source Separation. We use Spleeter, an deep learning based source separation tool developed by Deezer, to decompose each audio track into two stems: vocals and accompaniment (which includes instruments and background music). This separation enables a more fine-grained analysis of whether the protest signal is embedded more strongly in the lyrical delivery or in musical arrangement. For each stem, we extract CLAP and HuBERT embeddings and classify them independently to assess their contribution to protest prediction. Beyond individual stem analysis, we conduct a controlled mixing experiment: we combine the vocal tracks of protest songs with the accompaniment of non-protest songs and vice versa. This allows us to quantify which component vocal or instrumental carries more predictive weight in classification. We measure the percentage of mixed tracks classified as protest or non-protest, providing empirical insight into how each part contributes to the perception and modeling of protest music.

5 Results and Discussion

5.1 Text-based Results

Among the language models evaluated, XLM-RoBERTa achieved the highest performance with an F1-score of 91.10%, significantly outperforming both RoBERTa (82.66%) and DistilBERT (82.47%). Veucci's lyrics-to-genre model performed reasonably well with an F1-score of 80.82%, but still lagged behind the textual models including smaller ones , suggesting that linguistic features, rather than domain-specific lyric or musical cues, play

Model	Model Size	Accuracy	Precision	Recall	F1 Score
XLM-RoBERTa	270M	89.37%	89.26%	93.01%	91.10%
RoBERTa	125M	81.61%	83.41%	83.72%	82.66%
DistilBERT	66M	80.57%	84.52%	83.32%	82.47%
Ensemble	–	80.16%	82.04%	84.31%	81.64%
Veucci	110M	81.47%	84.97%	83.38%	80.82%
Logistic Regression	–	76.97%	75.24%	86.81%	80.61%

Table 2: Performance Comparison of Textual Models

a central role in distinguishing protest songs. To further explore this hypothesis, we trained logistic regression and ensemble models using only the extracted linguistic features. The linguistic features(along with p values (Fisher, 1925) used were: Average Line Length (p-value = 1.23×10^{-8}), Rhyme Density (p-value = 0.3928), Lexical Density (p-value = 3.13×10^{-4}), Sentiment Score (p-value = 4.26×10^{-8}), Unique Words (p-value = 7.76×10^{-4}), One-gram Repetition Rate (p-value = 4.06×10^{-16}), Two-gram Repetition Rate (p-value = 4.21×10^{-19}), Three-gram Repetition Rate (p-value = 1.08×10^{-18}) as shown in figure 1. Figure 1 illustrates clear

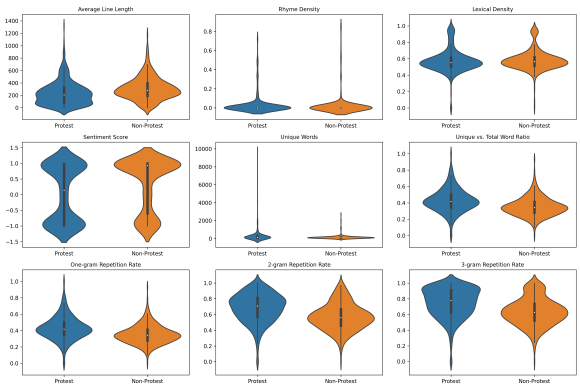

Figure 1: Comparison of linguistic features

linguistic distinctions between the two classes, particularly in n-gram repetition, sentiment scores, and lexical diversity—each significantly higher in protest songs. These models also outperformed Veucci, providing additional support for our claim. The results are displayed in Table 2. The statistical models were also trained and evaluated using the 80:20 split. This further strengthens our claim that in the textual dimension linguistic features are more significant than music specific lyrical features in distinguishing protest and non protest songs.

5.2 Audio-based Results

We evaluated three large-scale pretrained audio models CLAP, HuBERT, and Wav2Vec2 by extracting frozen embeddings and training lightweight classifiers on top of them. As shown in Table 3, CLAP significantly outperformed HuBERT and Wav2Vec2, achieving an F1-score of 90.62%. While CLAP is marginally larger in size, its superior performance is meaningful. Unlike HuBERT and Wav2Vec2, which are primarily trained on speech data, CLAP is trained to capture joint language-audio representations with a strong emphasis on music. It is thus more attuned to musical attributes such as timbre, rhythm, and expressive style. These results indicate that in the audio domain, music specific features not general acoustic or speech based cues play a more critical role in distinguishing protest songs from non-protest ones. In addition, we trained a logistic regression model on musical features extracted via Librosa, which achieved an F1-score of 86.45%. The Audio features used were `spectral_flatness` (9.30×10^{-25}), `spectral_flux` (1.28×10^{-21}), `mfcc` (7.73×10^{-17}), `rms` (1.39×10^{-16}), `repetition` (1.60×10^{-8}), `spectral_contrast` (1.70×10^{-6}) etc as shown in figure 2.

Despite its simplicity, this model outperformed both HuBERT and Wav2Vec2, reinforcing the insight that musically grounded features can outperform large models trained on general-purpose or speech-centric audio data. This further reinforces that in the audio

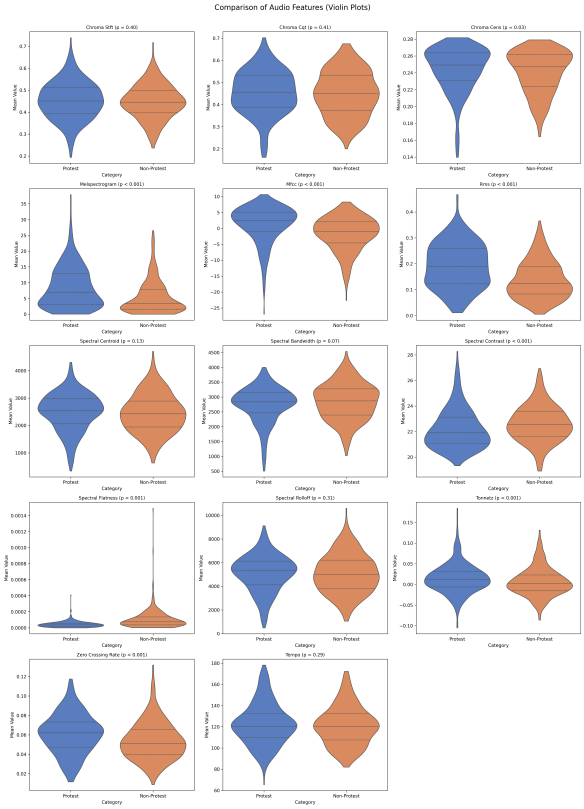

Figure 2: Comparison of audio features

domain, music-specific features are more effective than general-purpose or speech-based features in distinguishing between protest and non-protest songs.

5.3 Effect of Source Separation on Model Performance

As shown in Table 4, both CLAP and HuBERT achieved higher F1-scores for vocals (0.7470 and 0.6921, respectively) than for accompaniment (0.7273 and 0.6239). However, when evaluating mixed protest/non-protest tracks, both models attributed more protest content to the accompaniment. CLAP detected protest in 33.13% of accompaniment segments, compared to just 6.13% in vocals, while HuBERT flagged 65.64% of accompaniment and 42.33% of vocals. Despite HuBERT's overall higher protest detection rates, CLAP showed a smaller difference between vocal and accompaniment F1-scores (0.7470 vs. 0.7273), suggesting it relies more evenly on musical features. In contrast, HuBERT's higher protest detection in accompaniment could be due to its reliance on speech-like features, which may not generalize well to musical components. These results suggest that models may misattribute protest signals to accompaniment due to biases in how they interpret musical features, rather than reflecting a true distribution of protest cues between vocals and instrumentation.

5.4 Modality Comparison and Insights

Text-based models generally outperformed audio-based models in our dataset, particularly with larger pretrained transformers like XLM-R. However, the performance gap was not large: the best audio model (CLAP) was within 2–3% F1 of XLM-RoBERTa. This suggests that acoustic qualities such as vocal delivery, energy, and repetition are also strong indicators of protest intent. The competitive performance of interpretable linguistic features and statistical classifiers further supports the hypothesis that protest songs possess stylized, expressive cues that are detectable both textually and sonically.

5.5 Human Annotation Results

Nine musical and expressive features were annotated across protest and non-protest songs. Each feature was rated on a 5-point scale. The annotated features included perceived *speed* (tempo or pacing), *energy* (overall intensity, volume, and emotional charge), and *danceability* (rhythmic quality conducive to movement). We also evaluated *acousticness*, reflecting the degree of natural or acoustic instrumentation versus electronic sounds, and three dimensions of instrumentation: the complexity and presence of backing *instruments*, the prominence and clarity of *melody*, and the emphasis on *lyrics* in the mix. Additional features included *ornamentation*, referring to expressive musical flourishes such as trills, glides, and vibrato, and *disjunctness*, which measures melodic smoothness versus the presence of jumps or wide intervals. The results are summarized in Table 5, showing mean ratings for protest and non-protest songs, their differences, and the statistical significance (p-values) based on independent t-tests. The inter-annotator agreement test was conducted, we used cohen kappa (Cohen, 1960) for analyis, for all annotated musical features, and the results were as follows: Speed (Cohen's k =

Model	Size	Accuracy	Precision	Recall	F1-Score
CLAP	438M	**0.9130**	**0.9355**	**0.8788**	**0.9062**
HuBERT (Large)	317M	0.7938	0.7586	0.8327	0.7938
Wav2Vec2 (Large, 960h)	317M	0.6934	0.7000	0.6364	0.6666
Logistic Regression	–	0.8629	0.8655	0.8636	0.8645

Table 3: Performance of audio-based.

Model	Audio Type	Accuracy	Precision	Recall	F1 Score
HuBERT	Accompaniment	0.6985	0.7727	0.5231	0.6239
HuBERT	Vocal	0.7280	0.7321	0.6312	0.6921
CLAP	Accompaniment	0.7574	0.7857	0.6769	0.7273
CLAP	Vocal	0.7794	0.7600	0.7350	0.7470

Protest Component after mixing	CLAP (% Protest)	HuBERT (% Protest)
Vocals	6.13%	42.33%
Accompaniment	33.13%	65.64%

Table 4: Performance and protest detection rates of CLAP and HuBERT on source-separated audio.

0.58), Energy (Cohen's k = 0.54), Danceability (Cohen's k = 0.30), Acousticness (Cohen's k = 0.35), Disjunctness; melodic smoothness vs. jumps (Cohen's k = 0.30), Ornamentation; presence of extra musical effects (Cohen's k = 0.08), and Instrumentation Contribution: Melody (Cohen's k = 0.24), Lyrics (Cohen's k = 0.18), Instruments (Cohen's k = 0.28). These values indicate moderate agreement for Speed, Energy, Acousticness, and Instrumentation; Instruments, with fair to slight agreement for the rest. Since annotators did not have formal music training, lower consistency is understandable for more complex or technical features.

6 Conclusion

Our results reveal that protest music is primarily distinguished by general linguistic features rather than domain specific lyric or musical elements. Textually, the key differentiators are broad linguistic markers such as sentiment, lexical diversity, and n-gram repetition rate. These features suggest that protest songs rely on general linguistic cues that convey a sense of urgency, rebellion, or defiance, rather than on specific thematic or genre bound choices. In the audio domain, protest songs are more effectively characterized by music specific features. Notably, models trained on interpretable, genre agnostic features such as spectral flux and repetition from the Librosa library still achieved high scores. This reinforces that the observed patterns are not merely artifacts of genre. Through source separation and human evaluation, we observe that vocals play a more prominent role than accompaniment in distinguishing protest from non-protest songs. This aligns with the emotional intensity and rawness often associated with protest music. Yet, interestingly, our intermixing experiments reveal that accompaniment, while seemingly secondary, contributes more significantly than anticipated in shaping the perception of protest. The combination of instrumental and vocal elements particularly in how they interact appears to be a crucial factor in determining whether a song is perceived as protest music. Taken together, these findings suggest that protest music conveys its message through a multimodal approach: linguistically, by leveraging general textual signals that communicate the song's intent, and musically, by employing expressive and structurally distinct audio features. The interplay between these two domains text and music forms a holistic signature that makes protest music uniquely identifiable across both verbal and musical planes.

7 Future Work

This work lays the groundwork for understanding protest music as a multimodal vehicle of cultural resistance, aiming to explore its

Feature	Protest (Avg)	Non-Protest (Avg)	Difference	p-value
Speed	**3.97**	**2.17**	**1.80**	7.74×10^{-44}
Energy	**4.16**	**2.38**	**1.78**	2.71×10^{-41}
Danceability	**3.36**	**2.17**	**1.19**	4.84×10^{-13}
Acousticness	**2.03**	**3.40**	**-1.37**	1.59×10^{-19}
contribution of Instruments	**4.07**	**3.16**	**0.91**	1.13×10^{-9}
contribution of Melody	2.72	3.61	-0.89	5.69×10^{-8}
contribution of Lyrics	3.11	3.40	-0.29	0.107
Ornamentation (Musical Effects)	**3.46**	**3.07**	**0.40**	4.21×10^{-4}
Disjunctness (Melodic Jumps)	**3.32**	**2.22**	**1.10**	6.84×10^{-14}

Table 5: Human annotation results comparing protest and non-protest songs. Statistically significant differences ($p < 0.005$) (Dunn, 1961) after Bonferroni are in bold.

role in global social change. Future research can build upon this by expanding the dataset to include non-Western protest traditions such as Arabic *shaabi* and Korean *minjung kayo*, while also incorporating temporal metadata to facilitate diachronic and cross-cultural analysis. Although we aimed for genre balance during dataset construction, genre remains a potential confounding variable. Future studies should explicitly control for genre to ensure that observed distinctions are attributable to protest-related features rather than genre-specific conventions. On the modeling front, joint lyric-audio models with cross-modal attention offer a promising direction, particularly when fine-tuned on protest-specific corpora to better capture rhetorical nuance. Additionally, the growing influence of digital platforms warrants an investigation into how social media alters the creation, dissemination, and perception of protest music. Finally, incorporating human-centered evaluation such as listener surveys and focus groups will offer deeper insights into how protest intent is perceived by diverse audiences and can inform the design of more socially aware classification systems. To improve annotation consistency for complex musical features, future work may also consider involving trained musicians in the annotation process.

8 Ethical Considerations

All data used in this study, including song lyrics and audio excerpts, were obtained from publicly accessible, licensed platforms such as Spotify and YouTube, and analyzed strictly for academic research purposes under fair use provisions. The human annotation study was conducted with voluntary participants who were fully informed about the study's goals and procedures; no personal or identifiable information was collected. Throughout this project, we have remained attentive to issues of cultural sensitivity, particularly given the politically charged and historically grounded nature of protest music. Every effort was made to contextualize songs respectfully and accurately, avoiding reductive interpretations or cultural appropriation. Our goal is to amplify, not oversimplify, the expressive and political power of protest music across traditions and geographies.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Raffaella Bianchi. 2018. Istanbul sounding like revolution: The role of music in the gezi park occupy movement. *Popular Music*, 37:212–236.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Michael Conklin. 2014. Music and the civil rights movement. *Encyclopedia for Ethnomusicology*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995.

Support-vector networks. *Machine Learning*, 20(3):273–297.

D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Michael Drewett. 2003. *Music in the Struggle to End Apartheid: South Africa*, pages 153–165.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. Clap: Learning audio concepts from natural language supervision. *Preprint*, arXiv:2206.04769.

R. A. Fisher. 1925. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Yanru Jiang and Xin Jin. 2022. *Using k-Means Clustering to Classify Protest Songs Based on Conceptual and Descriptive Audio Features*, pages 291–304.

Amon Kipyegon Kirui. 2025. Music dualism: Political intolerance in kenya and the gen-z movement. *Journal of Music and Creative Arts*, 4(1):1–15.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25. Citeseer.

Holly Kingsley Miller. 1997. The times they were a-changing: A study of popular protest songs, 1963–1970. `https://digitalcommons.unomaha.edu/studentwork/2839`. Student Work, Paper 2839, University of Nebraska at Omaha.

OpenAI. 2023. Gpt-4 technical report. OpenAI Technical Report.

Lassane Ouedraogo. 2018. Pop music as e-civism: Negotiating change through subaltern voices in burkina faso. Presented at the Africana Studies Student Research Conference & Luncheon, Ohio University. Accessed online: https://ohioopen.library.ohio.edu/africana/2018/2.

Pratyay Raha. 2018. Role of music activism (ipta) in indian freedom movement – colonialism to a post-colonial context. In *MDW Book of Abstracts ISA 2018*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: Smaller, faster, cheaper, and lighter. *Preprint*, arXiv:1910.01108.

C. J. Van Rijsbergen. 1979. *Information Retrieval*, 2nd edition. Butterworths, London.