Project 3 Proposal

Vaishnavi Pradip Gayke
Binghamton University
Binghamton, New York, USA
vgayke@binghamton.edu

ABSTRACT

This project mostly aims to analyze social media discussions surrounding male and female music artists, focusing on identifying biases and patterns in sentiment, engagement, and toxicity. By utilizing data collected from Reddit's r/Music subreddit and 4Chan's /mu/ board, the project will explore how these platforms differ in their treatment of male and female artists, with a particular focus on sentiment trends, toxicity levels, and the influence of female artists' personal lives on public discourse. Through the development of an interactive tool, the project will present analyses to answer three key research questions, providing insights into gender-based perceptions in the music industry.

ACM Reference Format:
Vaishnavi Pradip Gayke. 2024. Project 3 Proposal. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social media platforms like Reddit and 4Chan have become influential arenas for discussing cultural and social topics, including music. These platforms reflect societal attitudes, often exposing biases in how male and female artists are perceived and critiqued.

Building upon the findings from Project 2, this project addresses the following research questions:

- How do social media platforms (Reddit and 4Chan) differ in the frequency of mentions and overall engagement for male vs. female music artists?
- What is the sentiment (positive vs. negative) surrounding male vs. female artists in online discussions, and how does toxicity manifest differently for each gender?
- What role does the personal life of female artists (e.g., relationships, controversies) play in shaping online discussions compared to male artists?

2 TOOLS AND TECHNOLOGIES

Here, we describe the libraries and tools that will be used throughout the project for data collection, cleaning, sentiment and toxicity analysis, statistical analysis, and visualization.

2.1 Libraries for Data Collection and Cleaning

- **Pandas**:
  - Used for data cleaning, manipulation, and preparation.
  - Provides powerful functions for handling structured data, such as filtering relevant records and transforming raw datasets into usable formats.
- **NumPy**:
  - Supports efficient numerical computations, such as aggregating engagement metrics and normalizing sentiment scores.
- **Requests**:
  - Simplifies API calls to collect data from Reddit's JSON API.
  - Used to interact with the *ModerateHatespeech API* for toxicity analysis.

2.2 Sentiment and Toxicity Analysis Libraries

- **Hugging Face Transformers**:
  - Offers pre-trained models like BERT, RoBERTa, and DistilBERT for advanced NLP tasks, including sentiment classification.
  - Provides better accuracy and contextual understanding compared to traditional libraries like TextBlob.
  - Particularly useful for nuanced text, such as sarcastic or complex statements.
- **TensorFlow/Keras**:
  - Allows building custom deep learning models for toxicity detection.
  - Enables fine-tuning models for specific datasets, which can outperform general-purpose APIs like ModerateHatespeech in certain contexts.

2.3 Libraries for Statistical and Comparative Analysis

- **SciPy**:
  - Enables statistical tests and hypothesis validation, such as comparing engagement metrics between male and female artists.
- **Statsmodels**:
  - Used for regression analysis to identify trends and patterns over time or across platforms.

2.4 Libraries for Visualization

- **Matplotlib**:
  - Provides a base for creating static visualizations like bar charts and time-series plots.
- **Seaborn**:
  - Built on Matplotlib for more aesthetically pleasing and high-level visualizations, such as heatmaps and distribution plots.
- **Plotly**:
  - Enables interactive visualizations with hover effects, zooming, and adjustable parameters.
  - Key for creating the interactive dashboard.
- **Wordcloud**:
  - Generates word clouds for qualitative analysis, such as visualizing common themes in toxic comments or event-based discussions.

2.5 Web Frameworks and Dashboards

- **Flask**:
  - A lightweight web framework for building the backend to serve dynamic content to the interactive dashboard.
- **Bootstrap**:
  - Ensures responsive design for the web interface, making it visually appealing and accessible on different devices.

3 ANALYSES

3.1 Platform Differences in Engagement

Compare the frequency of mentions and engagement metrics (e.g., upvotes, thread lengths) for male and female artists on Reddit and 4Chan.

- **Parameters**: Platform, artist, timeframe.
- **Visualization**: Bar charts and time-series graphs.

3.2 Sentiment and Toxicity Trends

Identify sentiment polarity (positive/negative) and toxicity differences for male and female artists across platforms.

- **Parameters**: Sentiment thresholds, platform filters.
- **Visualization**: Heatmaps and line graphs.

3.3 Impact of Female Artists' Personal Lives

Examine how personal life events of female artists influence online discourse compared to male artists.

- **Parameters**: Event-based filters, sentiment trends.
- **Visualization**: Trend graphs and word clouds.

4 INTERACTIVE TOOL

The interactive tool will include:

- Customizable filters (timeframes, platforms, artists).
- Interactive visualizations (hoverable graphs, dropdown filters).
- Export options (downloadable plots and summaries).

5 EXPECTED OUTCOMES

Key outcomes include:

- Platform-specific differences in sentiment, toxicity, and engagement.
- Insights into societal biases in the perception of female artists.

6 CONCLUSION

This project seeks to provide a comprehensive understanding of the biases in online music discussions about male and female artists. By integrating robust data analysis with interactivity, the findings aim to contribute to fairer digital and societal conversations.