Sexism and Violence Analysis in French Rap

AALABOU Mariem, ALLAL Alexia, SALAS Clara

PSL

November 29, 2024

Introduction

- Context: Importance of analyzing rap lyrics content
- Objective: Automated assessment of sexism and violence
- Approach: Comparison of two analysis methodologies

Corpus Overview

- Description of French rap corpus
- Time period covered
- Number of analyzed texts
- Selection criteria

Research Question

- How to automatically detect sexist and violent content in lyrics?
- Which approach is most effective?
- What are the advantages and limitations of each method?

Proposed Solutions

- Approach 1: Analysis using specialized LLMs
- Approach 2: Tokenization and lexical annotation
- Complementarity of approaches

Methodology - LLMs

Model Architecture

- XLM-RoBERTa for sexism detection
  - annahaz/xlm-roberta-base-misogyny-sexism
  - Specifications and parameters
- DehateBeRT for French
  - dehatebert-mono-french
  - Used configuration

LLM Explanation

Model Output

- **Logits:** `tensor([[-2.7563, 3.2781]])`
  - These are the raw scores before applying the softmax function.
  - They represent the model's confidence for each class.
- **Probabilities:** `tensor([[0.0024, 0.9976]])`
  - 0.24% for **"normal"** speech (class 0).
  - 99.76% for **"sexism/hate speech"** (class 1).
- **Predicted Class:** `1`
  - Here, the model predicts the text belongs to **class 1** (sexism/hate speech).

Methodology - Lexical Annotation

HurtLex Integration

- Categories for sexism detection:
  - PS (negative stereotypes), ASM/ASF (gendered terms)
  - PR (prostitution), OM (homosexuality)
- Categories for violence detection:
  - CDS (derogatory), SVP (deadly sins), RE (crime)
- Two-level structure:
  - Conservative: offensive senses only
  - Inclusive: all potentially relevant senses

Implementation Details

- Using HurtLex lexicon
- Text preprocessing: cleaning, normalization
- Batch processing with checkpointing
- Metrics calculation per 1000 songs

Lexical Annotation - Detailed Steps

Tokenization and Preprocessing

- Preprocess tokens: remove stop words, lemmatize, and deduplicate
- Split corpus into batches for parallel processing
- Use multiprocessing for efficient annotation
- Log progress and save intermediate results

Normalization of Rates

- Calculate sexism and violence rates
- Normalize rates as proportions of total tokens

Multiprocessing Method

Parallel Processing

- Utilized Python's `multiprocessing` library to parallelize the annotation process.
- Split the corpus into smaller batches to distribute the workload across multiple CPU cores.
- Each batch is processed independently, allowing for concurrent execution.
- Results are collected and concatenated after processing.

Results - Cartography of the Artists - LLMs

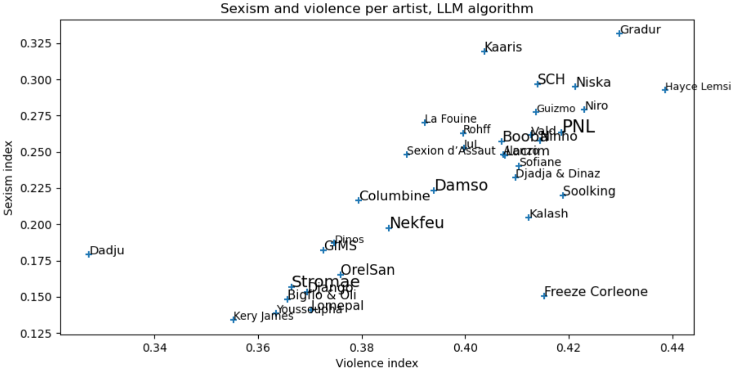


Figure: Cartography of the artists, using LLMs. X-axis is the mean violence index and y-axis is the mean sexism index of the artist, calculated on his 50 most popular songs. The name size represents popularity.

Results - Cartography of the Artists - Annotations

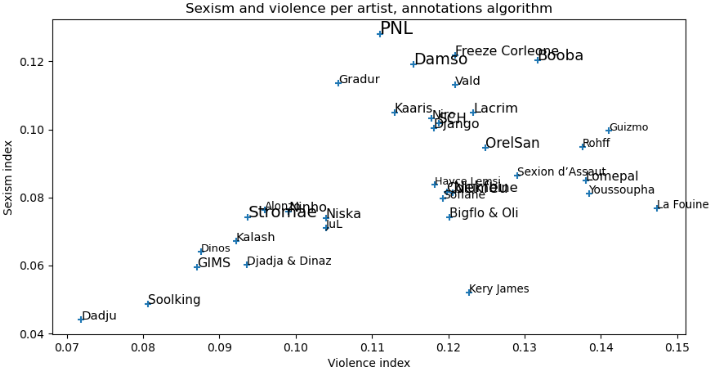


Figure: Cartography of the artists, using annotations. X-axis is the mean violence index and y-axis is the mean sexism index of the artist, calculated on his 50 most popular songs. The name size represents popularity.

Results - Temporal Evolution of Indexes - LLMs

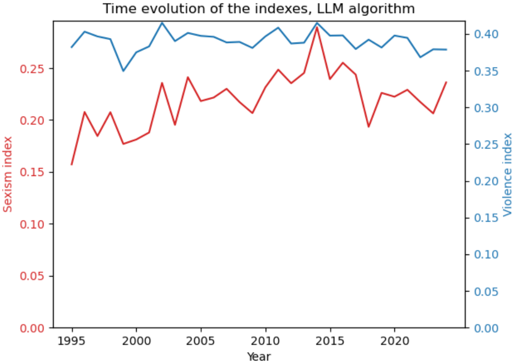


Figure: Temporal evolution of indexes. X-axis is the year and y-axis are the mean sexism and violence indexes, calculated on the 50 most popular songs of each year, using LLMs.

Results - Temporal Evolution of Indexes - Annotations

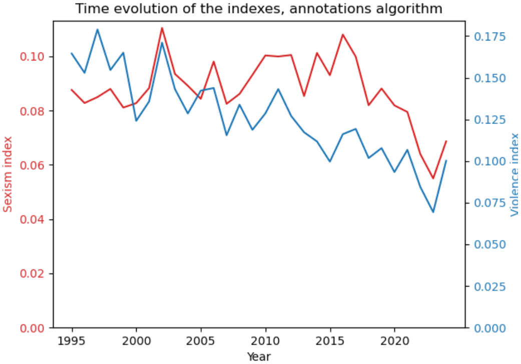


Figure: Temporal evolution of indexes. X-axis is the year and y-axis are the mean sexism and violence indexes, calculated on the 50 most popular songs of each year, using annotations.

Results - Temporal Evolution of Indexes per Artist - LLM

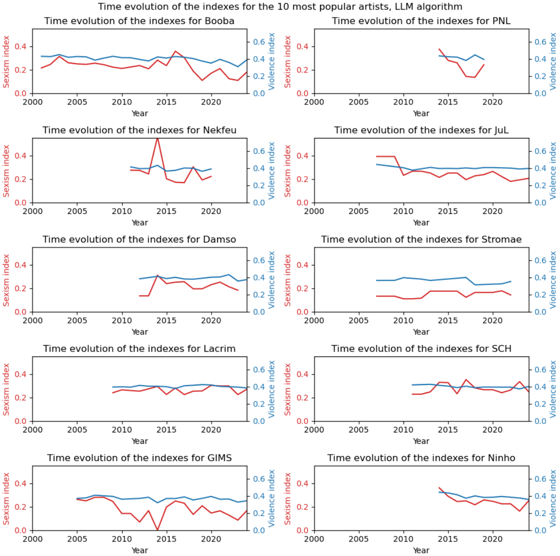
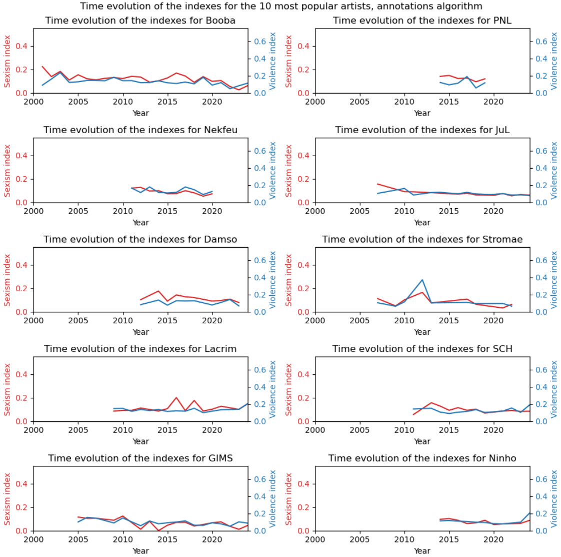


Figure: Temporal evolution of indexes for the 10 most popular artists. X-axis is the year and y-axis are the mean sexism and violence indexes, calculated on the 50 most popular songs of each year, using LLMs.

Results - Temporal Evolution of Indexes per Artist - Annotations


Time evolution of the indexes for the 10 most popular artists, annotations algorithm

Time evolution of the indexes for Booba

Time evolution of the indexes for PNL

Time evolution of the indexes for Nekfeu

Time evolution of the indexes for JuL

Time evolution of the indexes for Damso

Time evolution of the indexes for Stromae

Time evolution of the indexes for Lacrim

Time evolution of the indexes for SCH

Time evolution of the indexes for GIMS

Time evolution of the indexes for Ninho

Figure: Temporal evolution of indexes for the 10 most popular artists. X-axis is the year and y-axis are the mean sexism and violence indexes, calculated on the 50 most popular songs of each year, using annotations.

Conclusion

Summary

- This study presented a comparative analysis of two methodologies for detecting sexism and violence in French rap lyrics.
- The first approach utilized specialized Large Language Models (LLMs) such as XLM-RoBERTa and DehateBeRT.
- The second approach involved tokenization and lexical annotation using the HurtLex lexicon.
- Both approaches were evaluated for their effectiveness, advantages, and limitations.

Conclusion

Key Findings

- LLMs showed high accuracy in detecting sexism and violence but required significant computational resources.
- Lexical annotation provided detailed insights into specific terms and their contexts, but required extensive processing.

Limitations

- The LLMs used were pre-trained and may not capture nuances specific to French rap culture.
- The HurtLex lexicon, while comprehensive, may not include all relevant terms and contexts, potentially missing some instances of sexism and violence.
- The annotation process was time-consuming and required significant computational resources for batch processing.

Conclusion

Observations

- The detection of sexism and violence varied significantly among different artists and over time.
- There was a noticeable trend of decreasing sexism and violence indices in more recent years, suggesting a possible shift in the cultural norms within the genre.
- The annotation process highlighted the complexity and subjectivity involved in identifying sexist and violent content.

Future Work

Improvement Perspectives

- Future work could focus on optimizing the computational efficiency of LLMs.
- Enhancing the HurtLex lexicon with more context-specific terms could improve the accuracy of lexical annotation.
- Developing a hybrid model that integrates both LLMs and lexical annotation could provide more robust and accurate results.

References

Sources

- Used models:
  - annahaz/xlm-roberta-base-misogyny-sexism-indomain-mix-bal. Available at: https://huggingface.co/annahaz/xlm-roberta-base-misogyny-sexism-indomain-mix-bal
  - dehatebert-mono-french. Available at: https://huggingface.co/Hate-speech-CNERG/dehatebert-mono-french
- HurtLex documentation: https://github.com/valeriobasile/hurtlex/blob/master/lexica/FR/1.2/hurtlex_FR.tsv
- Relevant academic papers:
  - Courson, B. (2023). LRFAF: une exploration numrique du rap franais depuis les annes 1990.