Exploring Audience Reactions on YouTube: An Approach to Sentiment and Toxicity Analysis

André F. Rollwagen[1,2][a], Gabriel Zurawski[2][b], Stefano Carraro[2], Roberto Tietzmann[2][c], Marcelo C. Fontoura[2][d] and Isabel H. Manssour[2][e]

[1]*Federal Institute of Education, Science and Technology Sul-Rio-Grandense, IFSUL, Brazil*

[2]*Pontifical Catholic University of Rio Grande do Sul - PUCRS, Porto Alegre, Brazil*

Keywords: YouTube, Social Media Analysis, Sentiment Analysis, Toxic Speech Detection, Portuguese Language.

Abstract: Social media platforms like YouTube have become central spaces for public expression, allowing users to share their emotional reactions and, at times, toxic content. Understanding these reactions requires scalable and reproducible approaches that can handle informal, user-generated content. This paper presents an integrated approach for analyzing sentiment polarity and detecting toxic speech in YouTube video comments written in Portuguese. For this, we developed a set of Python scripts that automate data collection and apply Natural Language Processing (NLP) techniques to perform both tasks. These scripts are publicly available and can be adapted for use in various video and social analysis contexts. Interactive visualizations were also generated to support the interpretation of results. The applicability of the approach is demonstrated through two case studies involving highly controversial videos, which allow us to explore the relationship between sentiment, toxicity, and audience engagement patterns. The results provide valuable insights into the dynamics of public discourse and offer tools for future research on audience speech analysis on YouTube.

1 INTRODUCTION

Social media platforms like YouTube have become central spaces for public debate. Comments posted by users on videocasts or communication channels, e.g., often express opinions, emotions, and sometimes manifestations of toxic language. Although rich in content, such environments are also prone to offensive, discriminatory, or aggressive messages (Fortuna et al., 2021), reinforcing the need for automated methods to detect and interpret these expressions.

In this context, it is essential to clarify key concepts. Hate speech, as defined by (Fortuna et al., 2021), targets individuals or groups based on traits like race, gender, or religion, often carrying legal implications. Toxicity is a broader concept that includes hate speech along with offensive language, threats, and verbal abuse. Here, we use the term *toxic speech* to encompass this wider range of harmful expressions.

Most existing works focus exclusively either on sentiment analysis (Kurtz et al., 2025; Dharini et al., 2025) or on the detection of toxic speech (Bonetti et al., 2023; Maity et al., 2024). However, few works explore both dimensions jointly, especially in high-visibility contexts such as videocasts, where the combination of emotional tone and the presence of toxic speech can offer a more comprehensive understanding of audience reactions. Additionally, while much research centers on platforms like Twitter (Campan and Holtke, 2024; Siegel, 2020) (now called X), YouTube, despite its 2.7 billion users in 2023 (Muneer and Khan, 2025), remains underexplored. These figures underscore its strategic relevance for studies on large-scale interactions and discursive behavior.

Language bias is another limitation, with a predominance of English-centered analyses (Siegel, 2020). Portuguese remains underexplored, especially in works that provide open repositories and reproducible methodologies (Leite et al., 2020). In this context, we focus on the Portuguese language, motivated by both linguistic proximity and the nature of the conflicts analyzed, which emerged in Brazilian media and are expressed in that language.

Thus, this work aims to develop an approach

[a] https://orcid.org/0000-0002-1106-7222
[b] https://orcid.org/0009-0007-4594-1938
[c] https://orcid.org/0000-0002-8270-0865
[d] https://orcid.org/0000-0002-3229-0167
[e] https://orcid.org/0000-0001-9446-6757

Rollwagen, A. F., Zurawski, G., Carraro, S., Tietzmann, R., Fontoura, M. C. and Manssour, I. H.
Exploring Audience Reactions on YouTube: An Approach to Sentiment and Toxicity Analysis.
DOI: 10.5220/0013718200003985

In *Proceedings of the 21st International Conference on Web Information Systems and Technologies (WEBIST 2025)*, pages 481-488
ISBN: 978-989-758-772-6; ISSN: 2184-3252

for performing sentiment analysis and toxicity detection in Portuguese-language YouTube comments. The proposed solution involves the development of Python scripts that enable comment collection, sentiment classification (positive, negative, or neutral), and the identification of different levels of toxicity. Additionally, visualizations are generated to facilitate the interpretation of results and support the exploratory analysis of audience reaction patterns. To demonstrate the applicability of our approach, we analyzed two high-impact YouTube videos, examining how content type shapes the tone of comments and engagement patterns, supported by visualizations and metrics that reveal discursive trends.

Therefore, the main contributions of this work are:

- The integrated application of sentiment analysis and toxicity detection on the same textual dataset, enabling the investigation of relationships between emotional polarity and offensive speech.
- The set of Python scripts that automate the collection, sentiment analysis, and toxicity detection of comments on YouTube videos.
- The possibility of visually exploring correlations between sentiment and toxicity, allowing for the visual analysis of comment dynamics.
- The demonstration of the proposed approach through two case studies involving widely discussed videos, highlighting its ability to reveal discursive patterns in contexts marked by high audience engagement and social controversy.

The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 describes the methodology. The results are discussed in Section 4, together with the description of the two case studies. Section 5 outlines our findings.

2 RELATED WORK

Sentiment analysis and toxic speech detection on social media have been widely investigated as strategies to understand public opinion and engagement patterns in political, social, and cultural contexts. Recent advances apply Natural Language Processing (NLP) and deep learning techniques to automate the collection, preprocessing, and analysis of textual data (Yadollahi et al., 2022).

The study by (Parraga-Alava et al., 2021) employs an API-based pipeline to analyze tweet sentiment during Ecuador's presidential election, revealing correlations with the results. (Vargas et al., 2022) proposed an Unsupervised method for Aspect Term Extraction (UnATE) that combines topic models, word embeddings, and a fine-tuned BERT model, proving effective for sentiment analysis in product reviews without labeled data.

In YouTube, (Adesina and Howe, 2024) apply a multilabel model using Graph Convolutional Networks (GCNs) and LLMs to classify sentiments in political comments. Also, (Bindhumol et al., 2024) combine Convolutional LSTM and Convolutional Gated Recurrent Unit (CGRU) networks to support recommendation systems based on sentiments extracted from user comments. (Guzman et al., 2025) conduct a manual sentiment analysis to reveal patterns of distrust toward vaccines and public policies during the pandemic, while (Schmidt et al., 2023) analyze georeferenced sentiment on X after Elon Musk's acquisition, exposing regional polarization.

In toxic speech detection, (Shahi and Majchrzak, 2025) demonstrates that leveraging data from multiple platforms improves classifier performance when corpora are similar. (Kamma et al., 2025) propose a Bi-LSTM with hierarchical attention, achieving high precision and recall in comment classification.

Table 1 presents a comparative summary of related work, highlighting the data sources, models and libraries, and key contributions.

In contrast to related works, we propose an approach for analyzing YouTube comments written in Portuguese that combines sentiment analysis (positive, negative, and neutral) and toxic speech detection in a unified manner. While most works focus on English data analysis and typically address only one of these dimensions, we provide both classifications simultaneously. Interactive visualizations are also available to support the interpretation and exploration of the results. We made the code available on GitHub[1] to facilitate reproducibility and contribute to research on public speech in digital media.

3 RESEARCH METHODOLOGY

This section outlines the methodology used for collecting comments from YouTube videos, performing sentiment analysis, detecting toxic speech, and presenting the results through visualizations. To support this process, a set of Python scripts has been developed and is available, enabling the systematic execution of data gathering, processing, and analysis, the latter supported by LLMs. Figure 1 illustrates the methodological workflow adopted.

[1] https://github.com/DAVINTLAB/Toxicytube

Table 1: Comparison of related work.

Reference	Source	Sentiment Analysis	Toxic Detection	Visual Analysis	Language	Code Available	Main Contribution
(Parraga-Alava et al., 2021)	X	TextBlob [1]		✓	Spanish		Correlation between sentiment polarity and electoral outcomes
(Vargas et al., 2022)	Product Reviews	UNATE			English		Proposal of UNATE method for aspect-level sentiment analysis
(Adesina and Howe, 2024)	YouTube	GCNs, ChatGPT-4, VADER [2]		✓	English		Multilabel political sentiment classification using NLP
(Guzman et al., 2025)	X	Not specified		✓	Filipino, English		Patterns of distrust and negative sentiment during pandemic
(Bindhumol et al., 2024)	YouTube	CLSTM, CGRU		✓	English		Sentiment-based product recommendation from YouTube comments
(Schmidt et al., 2023)	X	RoBERTa		✓	English		Detection of negative sentiment and regional polarization
(Shahi and Majchrzak, 2025)	YouTube, X, Gab, Wikipedia		SVM, LSTM, BERT		English, German	✓	Improved hate speech detection using cross-platform dataset similarity
(Kamma et al., 2025)	X		Bi-LSTM, Word2Vec[3]		English		Hierarchical Bi-LSTM for toxic comment detection
Our Approach	YouTube	XLM-T, GPT-3.5-Turbo	XLM-T, GPT-3.5-Turbo, Detoxify	✓	Portuguese	✓	Combined analysis of sentiment, toxic speech detection, and visual exploration using dashboards

[1] TextBlob: Python library for processing textual data. [2] VADER: Valence Aware Dictionary and Sentiment Reasoner. [3] Word2Vec: Neural embedding technique that represents words as continuous vectors in a latent space, capturing their semantic similarity based on context.

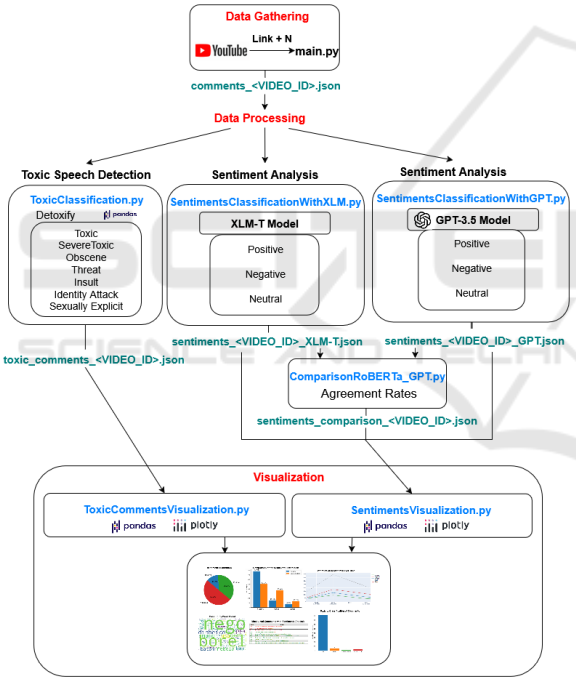

Figure 1: Methodological workflow adopted for the sentiment analysis and toxic speech detection.

3.1 Data Gathering

The data used in the proposed approach is obtained by collecting comments from videos published on YouTube, excluding live streams. To accomplish this, we developed a Python script (`main.py`) that uses the YouTube Data API[2] to perform comment collection from videos already available on the platform. To run the script, the user must provide a Google API key, the video ID, and the desired number (**N**) of comments. The system retrieves the top **N** most liked comments and saves them in a file named `comments_<VIDEO_ID>.json`, containing the fields *id*, *author*, *message*, *publishedAt*, and *likeCount*, which will be used in analysis steps.

[2]https://developers.google.com/youtube/v3

3.2 Data Processing

This section presents the automated processing of YouTube comments, combining pre-trained NLP models and GPT-3.5-Turbo for toxic speech detection and sentiment classification.

3.2.1 Toxic Speech Detection

Toxic speech detection is performed by the Python script `ToxicClassification.py`, using the JSON file generated during the data collection. The classification process uses the Detoxify (Hanu and Unitary team, 2020) library, which provides pre-trained transformer-based models to identify seven types of toxicity, including *toxicity* (general rude or disrespectful content), *severe_toxicity* (highly aggressive or hateful speech), *obscene* (use of offensive or profane language), *identity_attack* (targeting people based on attributes like race, gender, or religion), *insult* (personal attacks or name-calling), *threat* (statements expressing intent to harm), and *sexual_explicit* (sexual explicit language). The output is a `toxic_comments_<VIDEO_ID>.json` file that contains the original comment data along with the predicted scores for each toxicity type and serves as input for subsequent analyses.

A comment is assigned to a category if its predicted probability score exceeds the 0.7 threshold.

The Detoxify library was selected for its multilingual coverage and practical effectiveness in identifying toxic content on social media, including in Portuguese-language data.

3.2.2 Sentiment Analysis with XLM-T

Sentiment analysis is performed using two approaches. The first, via the Python script `SentimentsClassificationWithXLM.py`, applies the pre-trained model `cardiffnlp/twitter-xlm-roberta-base-sentiment`, fine-tuned for multilingual X data and based on the XLM-RoBERTa architecture (Barbieri et al., 2022).

It was chosen for its performance on short user-generated texts in the TweetEval benchmark (Barbieri et al., 2020). The model generalizes well to other languages, including Portuguese, despite being mainly trained on English data.

The script reads comments from a JSON file and classifies each as positive, negative, or neutral using the model. The predicted labels are added as a new column called `sentiment`, and the results are exported to `sentiments_<VIDEO_ID>_XLM-T.json`, indicating the model used for the classification.

3.2.3 Sentiment Analysis with GPT-3.5-Turbo

The second sentiment analysis approach was implemented using OpenAI's GPT-3.5-Turbo model via API, through the script `SentimentClassificationWithGPT.py`. GPT-3.5-Turbo was selected due to its high linguistic accuracy and contextual understanding, making it suitable for more nuanced sentiment analysis, particularly in user-generated content that often includes informal or ambiguous language (Zhang et al., 2024). However, its use requires access to the OpenAI API, which is subject to usage limitations and costs.

The script receives a JSON file containing user comments and applies a specific prompt designed to instruct the model to classify the sentiment of each comment, such as:

```
You are a sentiment analyzer.
Classify the sentiment of the following comment
as ''positive'', ''negative'', or ''neutral''.
Respond with the corresponding word only.
```

The result is exported to a JSON file named `sentiments_<VIDEO_ID>_GPT.json`, where the suffix `_GPT` indicates that the file was generated using the GPT-3.5-Turbo model.

3.2.4 Models Agreement

This step aimed to evaluate the consistency between the sentiment classifications produced by the XLM-T and GPT-3.5-Turbo models. To that end, we developed a script in Python named `ComparisonRoBERTa_GPT.py` ("RoBERTa" referring to the architecture underlying XLM-T). The input consists of two JSON files: one containing the sentiment labels assigned by XLM-T and the other by GPT-3.5-Turbo.

The script processes these files to calculate agreement rates between the models' classifications and generates summary statistics to assess alignment and potential divergences in interpretation. The output is a `sentiments_comparison_<VIDEO_ID>.json` file containing only the comments for which both models assigned the same sentiment rating.

3.3 Results Visualization

To support the interpretation of the classified data, we developed two Python scripts: `SentimentsVisualization.py`, for visualizing sentiment classification results, and `ToxicCommentsVisualization.py`, for visualizing toxic speech detection results. Both use the `Plotly` library[3] to generate interactive visualizations based on their respective JSON input files.

The script generates several visualizations, including time series plots that display temporal variations in sentiment polarity or toxicity categories, bar charts showing the total number of comments per sentiment or toxicity label, pie charts representing these proportions as percentages, and word clouds to highlight the most frequent terms. The resulting visualizations are organized into dashboards and exported as interactive HTML files, as shown in Figure 2.

4 RESULTS AND DISCUSSION

This section presents and discusses the results of applying our methodology to two YouTube case studies: the Monark case [4] (Section 4.1) and the Nego do Borel case [5] (Section 4.2). Section 4.3 compares both cases. The analysis addresses sentiment, toxicity, and user engagement patterns.

The selected videos were chosen for their high visibility and controversial nature, making them suitable for analyzing patterns of sentiment and toxic speech. The Monark case centers on a politically charged statement about Nazism, which sparked public outrage and moral condemnation. In contrast, the Nego

[3]https://plotly.com/python/

[4]https://www.youtube.com/watch?v=Qo2kYS2_XnI

[5]https://www.youtube.com/watch?v=FY3m6hMyh3g

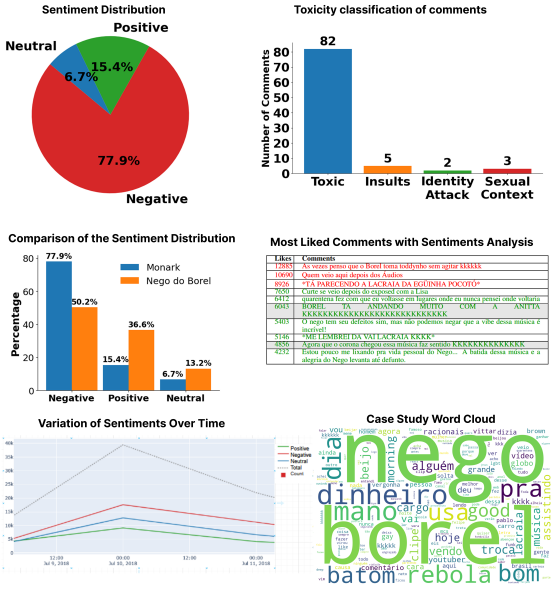

Figure 2: Interactive visualizations in HTML format.

do Borel case revolves around a music video that generated diverse reactions, influenced by cultural references and the artist's public image. Both were selected for their thematic relevance, high engagement, and suitability for contrasting public responses.

We defined three hypotheses based on the themes and public impact of the videos. First, we expected a peak in comment activity for the Monark case shortly after the video's release, due to the controversy's intensity. Second, we assumed that political polarization could intensify user engagement, leading users to actively seek out related content. Lastly, we hypothesized that top-liked comments, while critical, would avoid overtly toxic or hateful language, favoring provocative but moderate discourse. These hypotheses are examined in Sections 4.1, 4.2, and 4.3.

The comments were collected from each video's publication date until May 7, 2025. For the Monark case, the collection period began on February 9, 2022, yielding 3,798 comments. In the Nego do Borel case, it started on July 9, 2018, with a total of 114,695 comments. It is important to note that the script used does not collect comment replies.

To ensure a fair comparison despite the difference in comment volume, we selected a sample of the 1,000 most liked comments per case study. This sampling enabled a balanced analysis in terms of volume, focusing on the most relevant and highly engaged content.

We applied two models for sentiment classification: GPT-3.5-Turbo (via prompts) and XLM-T (multilingual, fine-tuned). Both models were applied to the datasets in parallel, allowing us to compare their outputs and assess their consistency. This model agreement step was applied to both case studies, with 77.2% agreement for Monark and 56.9% for Nego do Borel. Due to the low alignment between the models, we based our analyses on GPT classifications, given its strong zero-shot and few-shot performance, as well as its previously discussed advantages (Zhang et al., 2024), requiring no additional training.

4.1 Monark Case

This case study refers to the audience reaction to controversial statements by Bruno Aiub (Monark) on the Flow Podcast [6]. As the original video was removed from the platform, making direct analysis unfeasible, we used an alternative source, a video published by the Jornalismo TV Cultura channel, which reported his dismissal from the program under the headline "Monark é demitido, após defesa de partido nazista" (translated as "Monark is fired after defending the existence of a Nazi party"). The removed video, which lasted 1 minute and 46 seconds, had garnered 1,414,319 views and 6,081 comments.

We analyzed sentiment and toxicity to assess public reaction. As shown in Figure 3a, the results indicate a strong prevalence of criticism, with 77.9% of the comments classified as negative. Although the video had a journalistic focus, users used the comment section to voice direct criticism and moral judgment.

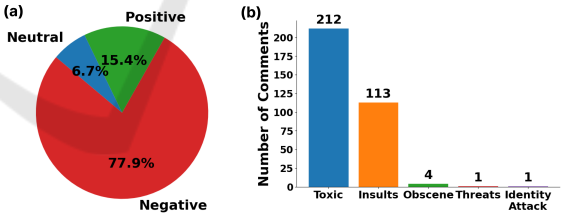

Figure 3: Classification of audience reactions by sentiment and toxicity Monark case.

Just 15.4% of comments conveyed positive sentiment, reflecting limited support, irony, or partial defense. In the sample, 33% of the comments showed some degree of toxicity: 21.2% were toxic and 11.3% insults (212 and 113 comments, respectively), as shown in Figure 3b). These results indicate that, although criticism prevailed, a significant portion of the engagement occurred through offensive language. More severe categories like threats and identity attacks were rare, supporting the hypothesis that top-

[6]https://www.youtube.com/@FlowPodcast.

liked comments, though critical, avoid extreme or explicitly hateful speech.

Figure 4 shows the most frequent terms in the comments, providing an overview of dominant themes and expressions used by the audience.

Figure 4: Audience reactions to the Monark case.

Daily comment volume was analyzed to understand engagement over time. Activity peaked on the first day (391 comments), then declined rapidly, indicating a brief but intense reaction typical of viral controversies, with limited sustained discussion.

To support contextualization, Table 2 presents the top 10 most liked comments that reveal a discursive plurality, combining direct criticism and political reflections. While some contain toxic elements, they are not dominant. The most liked comment — "Monark é a melhor propaganda anti drogas q existe kkkkjkkk" (translated as "Monark is the best anti-drug advertisement that exists LOL") — classified as positive, exemplifies the use of humor as an indirect criticism.

Table 2: Partial list with top 10 most liked comments.

Likes	Comments
5590	Monark é a melhor propaganda anti drogas q existe kkkkjkkk
5504	O maior erro dele foi misturar lazer com coisa séria. Se drogar num debate político é o cúmulo da irresponsabilidade.
2219	Esse cara é o espécime perfeito para oq as pessoas chamam de idiota
1733	O professor que explicou sobre o holocausto foi bem assertivo: QUANDO VOCÊ É TOLERANTE COM O INTOLERÁVEL...
1665	Monark acha que liberdade de expressão é nao ter responsabilidade. Cara defender que alguem seja antissemita racista . . .
1504	Perdeu dinheiro e se arrependeu, se não fosse isso...
1336	Bem que diziam que um dia, o Monark ia falar uma porcaria MUITO grande
1267	Bons tempos em que monark era apenas uma marca de bicicletas. Saudades.
870	O cara foi demitido não por causa do que ele falou e sim pela perdas dos patrocinadores. O mundo gira em torno do dinh. . .
603	Se esse cara vai trabalhar bêbado não tinham que trabalhar com ele. É uma irresponsabilidade

Negative sentiments are shown in red, positive in green, and toxic comments are highlighted with a light gray background.

The findings indicate a predominance of negative sentiment and offensive discourse. However, the low presence of severe toxicity among top-liked comments suggests a social moderation, with less approval for more aggressive content.

4.2 Nego Do Borel Case

This case study examines the audience reception of the music video "Me Solta" by Nego do Borel. At the time of data collection, the 3 minutes and 46 seconds video, published on YouTube, had over 231 million views and 114.695 comments, including the replies.

Figure 5a depicts the proportions of positive, neutral, and negative comments, while Figure 5b shows the frequency of toxic content. The sentiment analysis reveals that over half of the comments were negative. Despite being entertainment content, many users used the comment section to express criticism or controversy toward the artist or the music video. Toxicity analysis revealed a low incidence of harmful speech: only 8.2% of comments were classified as toxic, and less than 1% into severe categories.

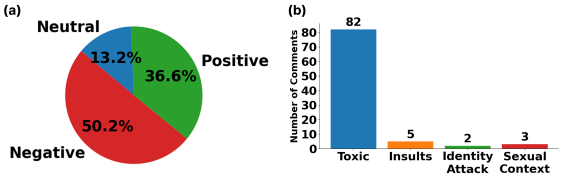

Figure 5: Classification of audience reactions by sentiment and toxicity in the Nego do Borel case.

A word cloud showing audience reactions to the video, displayed in Figure 2, highlights the most frequently used terms and offers qualitative insights into the main themes and expressions present in public discourse.

To analyze engagement in the Nego do Borel case, we examined comment distribution over time. The peak occurred on July 10, 2018 (270 comments), one day after publication, followed by a gradual decline - 173 and 71 comments on the two subsequent days. This pattern reflects typical viral behavior: intense initial reaction followed by rapid drop-off.

The table with the most liked comments presented in Figure 2 helps illustrate the tone and framing adopted by the audience, often marked by informal language and cultural references. For instance, "ME LEMBREI DA VAI LACRAIA KKKK" (translated as "REMINDED ME OF VAI LACRAIA LOL") and "Borel tá andando muito com a Anitta KKKK" ("Borel is hanging out with Anitta too much LOL") reference pop culture figures, use humor and sarcasm.

Overall, the Nego do Borel case suggests that audience engagement was driven more by cultural references and entertainment than by moral judgment or ideological polarization.

4.3 Comparative Analysis

The comparison between the Monark and Nego do Borel cases reveals distinct patterns in public reception, sentiment, and toxicity. While both videos generated high engagement, the tone and nature of com-

ments reflect their differing sociopolitical and cultural contexts. Figure 6 summarizes these contrasts, showing sentiment polarity (6a) and toxicity levels (6b) for each case.

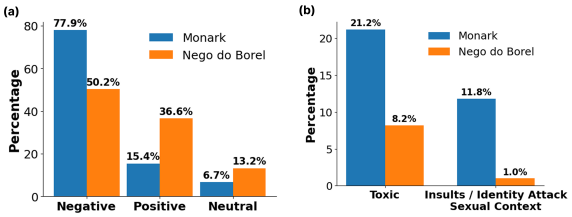

Figure 6: Sentiment and toxicity distribution in case studies.

In the Monark case, 78% of comments were negative, reflecting moral and political criticism, and turning the comment section into a space for public accountability and ideological positioning. In contrast, the Nego do Borel case showed a more balanced sentiment, with lighter, pop culture–driven reactions despite a high rate of negative comments (50%).

Regarding toxicity, the Monark case exhibited a higher incidence of toxic speech, though severe categories were rare, indicating audience tolerance for criticism within certain limits. In contrast, toxicity in the Nego do Borel case was lower and milder, often expressed through humor, irony, or nostalgia.

The nature and context of each case shaped audience responses: the Monark case, rooted in political controversy, triggered moral and hostile reactions, while the Nego do Borel case, framed as entertainment, prompted lighter, culturally driven commentary. The comparison shows that audience engagement depends not only on controversy severity but also on how the audience frames the content. While the Monark case sparked explicit outrage, the Nego do Borel case elicited more ambiguous and emotionally driven responses, with less toxicity. These findings suggest that sentiment polarity and toxicity are shaped by content themes and social framing.

The initial hypothesis assumed intense comment activity in the Monark case immediately after the video's release, which the data confirmed. For the Nego do Borel video, we anticipated longer-term engagement, but it also peaked early, with a significant majority of comments concentrated within the initial days. These findings suggest that the viral dynamics of YouTube engagement tend to follow an intense but short-lived cycle, regardless of content type.

In the Monark case, many comments appeared on a journalistic video. This supports our second hypothesis: that political polarization drives users to engage with related content even after its removal. It is plausible that many users had watched the original video or excerpts and used the comment section to react to the controversy.

5 CONCLUSIONS

This work proposed an integrated and reproducible approach to analyze sentiment and detect toxic speech in Portuguese-language YouTube comments. By developing a set of Python scripts, we automated the processes of comment collection, sentiment classification using both GPT-3.5-Turbo and XLM-T, and toxicity detection using the Detoxify library. The generated outputs, complemented by interactive visualizations, provided a comprehensive view of audience reactions to controversial content. We have made the source code of our approach openly available on GitHub for anyone who wants to use it, along with the case studies, including data and results.

The two case studies revealed distinct patterns of audience response. The Monark case demonstrated a predominance of negative sentiment and toxic language, underscoring the impact of political polarization and moral judgment. The Nego do Borel case exhibited a more heterogeneous emotional distribution, with pop culture references contributing to a discourse that was less aggressive overall. These findings suggest that public sentiment and toxicity are shaped not only by the content itself but also by how the audience socially frames the message and the broader context in which the communication occurs.

As future work, we plan to expand the scope to multilingual datasets and other social platforms. We also aim to enhance the interpretability of classification outcomes by utilizing explainable AI techniques, thereby contributing to greater transparency and accountability in the analysis of digital public discourse.

ACKNOWLEDGEMENTS

Carraro would like to thank the Tutorial Education Program (PET). Manssour would like to thank the financial support of the CNPq Scholarship - Brazil (303208/2023-6). This paper was supported by the Ministry of Science, Technology, and Innovations, with resources from Law No. 8.248, dated October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex, and published in the Residência em TIC 02 - Aditivo, Official Gazette 01245.012095/2020-56.

While preparing and revising this manuscript, we used ChatGPT, Grammarly, and Google Translate to enhance clarity and grammatical precision, as English

is not our first language. The authors take full responsibility for the content and its technical accuracy.

REFERENCES

Adesina, M. T. and Howe, L. (2024). Social media (youtube) political sentiment multi-label analysis. *International Journal of Science and Research Archive*, 12(02):2063–2071.

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., and Ronzano, F. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1644–1650.

Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266.

Bindhumol, M., Singh, T., and Patra, P. (2024). Sentiment analysis using youtube comments. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Bonetti, A., Martínez-Sober, M., Torres, J. C., Vega, J. M., Pellerin, S., and Vila-Francés, J. (2023). Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. *Applied Sciences*, 13(10).

Campan, A. and Holtke, N. (2024). Beyond twitter: Exploring alternative api sources for social media analytics. In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024), Vol.1*, pages 441–447.

Dharini, N., Madhuvanthi, M., Aswini, C. S., Lakshya, R., Triumbika, M., and Saranya, N. (2025). Ensemble-driven multilingual sentiment analysis framework for youtube comments with dashboard. In *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*, pages 1524–1529.

Fortuna, P., Soler-Company, J., and Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Guzman, J. P. S., Cheng, C. K., Bernadas, J. M. A. C., and Lao, A. R. (2025). Aspect-level sentiment analysis of filipino tweets during the covid-19 pandemic. In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Vol.2*, pages 343–350.

Hanu, L. and Unitary team (2020). Detoxify. Github. https://github.com/unitaryai/detoxify.

Kamma, V., Wbaid, S., K, S., A, N., and Saravanan, T. (2025). Toxic comment detection based on improved bi directional long short term memory. In *2025 International Conference on Intelligent Systems and Computational Networks (ICISCN)*, pages 1–6.

Kurtz, G. B., de P. Carraro, S., Teixeira, C. R. G., Bandeira, L. D., Müller, B. L., Tietzmann, R., Silveira, M. S., and Manssour, I. H. (2025). Streamvis: An analysis platform for youtube live chat audience interaction, trends and controversial topics. In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Vol.2*, pages 630–640.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.

Maity, A., More, R., Patil, P. A., Oza, J., and Kambli, G. (2024). Toxic comment detection using bidirectional sequence classifiers. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 709–716.

Muneer, Q. and Khan, M. A. (2025). Role of youtube in creating awareness of sustainable transportation: A latent dirichlet allocation approach. *Sustainable Futures*, 9:100607.

Parraga-Alava, J., Rodas-Silva, J., Quimi, I., and Alcivar-Cevallos, R. (2021). Opinion and sentiment analysis of twitter users during the 2021 ecuador presidential election. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies (WEBIST 2021)*, pages 257–266.

Schmidt, S., Zorenböhmer, C., Arifi, D., and Resch, B. (2023). Polarity-based sentiment analysis of georeferenced tweets related to the 2022 twitter acquisition. *Information*, 14(2):71.

Shahi, G. K. and Majchrzak, T. A. (2025). Hate speech detection using cross-platform social media data in english and german language. In *Proceedings of the 20th International Conference on Web Information Systems and Technologies (WEBIST 2024)*, pages 131–140.

Siegel, A. A. (2020). *Online Hate Speech*, pages 56–88. SSRC Anxieties of Democracy. Cambridge University Press, Cambridge.

Vargas, D. S., Pessutto, L. R. C., and Moreira, V. P. (2022). Unsupervised aspect term extraction for sentiment analysis through automatic labeling. In *Proceedings of the 18th International Conference on Web Information Systems and Technologies (WEBIST 2022)*, pages 344–354.

Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2022). Deep learning in sentiment analysis: Recent architectures and trends. *ACM Computing Surveys*, 54(2):1–36.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., and Bing, L. (2024). Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3883–3906.