UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Informatics

Sexism Identification on TikTok: A Multimodal AI Approach with Text, Audio, and Video

End of Degree Project

Bachelor's Degree in Data Science

AUTHOR: Arcos Gabaldon, Ivan

Tutor: Rosso, Paolo

Experimental director: Chulvi Ferriols, María Alberta

ACADEMIC YEAR: 2023/2024

Summary

Sexism persists as a pervasive issue in society, particularly evident on social media platforms like TikTok. This phenomenon encompasses a spectrum of expressions, ranging from subtle biases to explicit misogyny, posing unique challenges for detection and analysis. While previous research has predominantly focused on textual analysis, the dynamic nature of TikTok demands a more comprehensive approach. This study leverages advancements in Artificial Intelligence (AI), specifically multimodal deep learning, to establish a robust framework for identifying and interpreting sexism on TikTok. We compiled the first dataset of TikTok videos tailored for analyzing sexism in both English and Spanish. This dataset not only provides a foundational resource for current analysis but also serves as an initial benchmark for comparing models or for future investigations in this area. By integrating text, linguistic features, emotions, audio, and video features, this study identifies unique indicators of sexist content. Multimodal analysis surpasses text-only methods, particularly in understanding the intentions behind sexism, achieving remarkable results with F1-macro scores of 0.753 and 0.768 for English and Spanish, respectively. Notably, this configuration led to an improvement of 4.4% and 4.8% over the best unimodal models. Further, employing fine-tuning to a multimodal model (TAVL - Fine-Tuning), the results improve for all tasks, with a 5.5% increase in F1-macro for detecting sexism in English and a 2.2% improvement in Spanish. Additionally, for source intention, the improvements are 7.3% and 9.4%, respectively. Notably, for categories of sexism, there is a significant enhancement particularly in Spanish, where the categories are better represented and there are more sexist videos than in English.

Keywords – Multimodal Sexism Identification, TikTok, Artificial Intelligence.

Resumen

El sexismo persiste como un problema generalizado en la sociedad, particularmente evidente en plataformas de redes sociales como TikTok. Este fenómeno abarca un espectro de expresiones, que van desde sesgos sutiles hasta misoginia explícita, planteando desafíos únicos para su detección y análisis. Mientras que investigaciones previas se han centrado predominantemente en el análisis textual, la naturaleza dinámica de TikTok exige un enfoque más integral. Este estudio aprovecha los avances en Inteligencia Artificial (IA), específicamente el aprendizaje profundo multimodal, para establecer un marco robusto para identificar e interpretar el sexismo en TikTok. Compilamos el primer conjunto de datos de videos de TikTok diseñados para analizar el sexismo tanto en inglés como en español. Este conjunto de datos no solo proporciona un recurso fundamental para el análisis actual, sino que también sirve como un referente inicial para comparar modelos o para futuras investigaciones en esta área. Integrando texto, características lingüísticas, emociones, audio y características de video, este estudio identifica indicadores únicos de contenido sexista. El análisis multimodal supera los métodos solo textuales, particularmente en la comprensión de las intenciones detrás del sexismo, logrando resultados notables con puntajes F1-macro de 0.753 y 0.768 para inglés y español, respectivamente. Notablemente, esta configuración llevó a una mejora del 4.4% y 4.8% sobre los mejores modelos unimodales. Además, empleando el ajuste fino a un modelo multimodal (TAVL - Fine-Tuning), los resultados mejoran para todas las tareas, con un aumento del 5.5% en F1-macro para detectar sexismo en inglés y una mejora del 2.2% en español. Adicionalmente, para la intención de fuente, las mejoras son del 7.3% y 9.4%, respectivamente. Notablemente, para las categorías de sexismo, hay un mejoramiento significativo particularmente en español, donde las categorías están mejor representadas y hay más videos sexistas que en inglés.

Keywords – Identificación Multimodal de Sexismo, TikTok, Inteligencia Artificial.

Acknowledgements

I am deeply grateful to my parents, whose love and guidance are my constant inspiration. Their unwavering support has been my bedrock throughout this journey.

My heartfelt thanks go to my friends, especially Jaime and Pablo. Their companionship and encouragement have not only made this academic journey enjoyable but have also been a pillar of my strength during challenging times.

I would like to extend my sincere gratitude to my supervising professor, Paolo Rosso, for his invaluable trust and guidance. His close involvement and insightful feedback have been crucial in shaping both the direction and the success of this project.

A special thank you to Berta Chulvi for her organizational skills and dedication, which were instrumental in the creation of the dataset used in this research. Each of you has played a pivotal role in my academic and personal growth, and I am eternally grateful for your contributions to my life and studies.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Sexism and Social Media

Sexism refers to multifaceted, encompassing subtle expressions that can be as insidious as explicit misogyny. Whether presented as seemingly positive remarks, jokes, or offensive comments, sexism permeates various aspects of individuals' lives, influencing domestic and parenting roles, career opportunities, sexual image, and life expectations. Recognizing the diverse forms of sexism is crucial to understanding its impact on society. Glick and Fiske distinguish between two forms of sexism: hostile and benevolent sexism [16]. Hostile sexism includes openly negative and antagonistic attitudes toward women, whereas benevolent sexism is characterized by subjectively positive attitudes that are nevertheless tainted with chivalry, implying a protective stance that reinforces traditional gender roles.

Social media platforms have become conduits for the dissemination of sexist content, perpetuating and even normalizing gender differences and biased attitudes. The Internet, with its vast reach, reflects and amplifies societal inequalities and discrimination against women. This study is particularly crucial given the significant presence of teenagers on social media platforms, urging the need for urgent investigation and societal dialogue, especially from an educational standpoint.

TikTok, a dynamic social media platform with 1.218 billion users aged 18 and above, has revolutionized content consumption. Known for its role in shaping fast-paced, short-form videos, TikTok has become a hub for diverse content creation and dissemination. Out of 5.3 billion internet users worldwide, 23% actively engage with TikTok[1]. A study conducted by The Observer revealed that TikTok's algorithm can lead users down a path of increasingly sexist content[2]. This raises concerns about the platform's potential to reinforce negative preconceptions and misogyny. If someone comes to the app already thinking negatively about a group, TikTok's algorithm shows them more content that supports and even makes those negative thoughts seem okay.

1.2 Motivation

TikTok's danger in spreading sexist content is pronounced due to its rapid content turnover and the sheer volume of daily uploads. The platform's algorithm faces a challenge in effectively preventing this content, exacerbated by the fact that sexism is often camouflaged and difficult to detect.

Adding to the concern, recent research from the University of Portsmouth[3], highlights TikTok's role in amplifying extremist ideologies, particularly among groups like incels who propagate misogyny, sexism, and even violence against women. The study concludes that TikTok serves as a platform for spreading content that encourages violence against women. This underscores the urgency to reassess TikTok's content moderation strategies, considering the difficulty in detecting disguised sexism, to address its contribution to the dissemination of harmful beliefs.

Moreover, the motivation extends to the multidimensional nature of sexism on TikTok. While previous Natural Language Processing (NLP) studies have focused on textual analysis to detect sexist content, the incorporation of audio and video elements on TikTok demands a more comprehensive approach. The

[1] https://www.businessofapps.com/data/tiktok-report/

[2] https://medium.com/moviente/does-tiktok-have-a-misogyny-problem-c1033fbb2cc2

[3] https://www.port.ac.uk/news-events-and-blogs/news/new-research-highlights-the-role-of-tiktok-in-spreading-videos-that-encourage-violence-against-women

existing advancements in Artificial Intelligence (AI), particularly in deep learning, have primarily operated within single modalities, such as text or image recognition. However, the emergence of multimodal AI presents an exciting frontier in research. By combining text, audio, and video analysis, we aim to create a more nuanced and robust system for detecting and understanding sexist content on TikTok. This multidimensional approach aligns with the evolving capabilities of AI, offering the potential to enhance society by addressing the challenges posed by the dynamic nature of content on social media platforms.

1.3 Objectives and Research Questions

In the rapidly evolving landscape of social media, the issue of sexism persists and takes new forms, particularly on platforms like TikTok. The dynamic nature of content, rapid turnover, and the sheer volume of daily uploads pose significant challenges in effectively detecting and addressing sexist content on this platform. This research aims to explore and understand sexism on TikTok through a multidimensional lens, considering the incorporation of text, audio, and video elements in content creation. In the framework of this Final Degree Project we aim at answering the following Research Questions:

- **RQ1.** What features distinguish sexist TikTok content from non-sexist content, and how do these features contribute to identifying the source intention and categorization of sexism on the platform?
- **RQ2.** How well does GPT-3.5 Turbo perform in annotating sexist content on TikTok compared to human annotators, and what is the level of agreement between GPT-3.5 Turbo and human annotators?
- **RQ3.** How effective are classifiers based on single modalities (text, audio, and video) in detecting sexism, determining source intention, and categorizing different forms of sexism on TikTok? This question seeks to evaluate the individual strengths and limitations of classifiers focusing on specific modalities.
- **RQ4.** Do classifiers utilizing a multimodal approach, combining text, audio, and video analysis, outperform single modality classifiers in terms of detecting sexism, understanding source intention, and categorizing different manifestations of sexism on TikTok?

This research aims to contribute to the ongoing discourse surrounding the impact of social media on perpetuating sexist content, especially on platforms like TikTok. By adopting a multidimensional approach and leveraging advancements in AI, the study aims to provide insights into the detection and understanding of sexism in a diverse and dynamic content environment.

1.4 Thesis Structure

The thesis is structured into six chapters. Chapter 2 introduces related work, providing an overview of existing literature and methodologies, including textual and multimodal work, and addressing legal and ethical considerations. Chapter 3 details the dataset, the Servipoli annotation process, and tasks related to detecting and categorizing sexism in TikTok videos, with summaries of Spanish and English TikTok corpora, and includes Cohen's Kappa agreement scores across different tasks. Chapter 4 covers the preprocessing and feature extraction for text, audio, video, and multimodal models, discussing various linguistic resources, feature representations, and model architectures. Chapter 5 presents the experimental setup, metrics, and results for text, audio, video, and multimodal experiments, highlighting feature importance and model performance across different tasks. Chapter 6 concludes with a summary of findings, implications, and suggestions for future research directions.

Chapter 2

Related Work

2.1 Introduction

Hate Speech (HS) is generally described as any form of communication that belittles a person or a group based on attributes such as race, ethnicity, gender, sexual orientation, nationality, religion, among others [22]. When the target of hate speech is women, it manifests as a form of misogyny. However, *misogyny*, as defined by the Oxford English Dictionary[1], refers to feelings of hatred or dislike towards women, or beliefs that devalue women compared to men. Misogyny can exist in behaviors, attitudes, or beliefs that demean women or see them as inferior to men, without the need for overt hate speech. On the other hand, *sexism* is defined as prejudice, stereotyping, or discrimination, often against women, based on sex. Unlike misogyny, sexism can manifest subtly, such as through gender stereotypes, traditional gender roles or unequal access to opportunities[2].

In recent years, the proliferation of social media has amplified the visibility and spread of hate speech, misogyny, and sexism. This increase poses significant challenges but also highlights the necessity for developing automated tools to aid in detecting and mitigating such harmful content. Achieving fully autonomous systems for this purpose is complex due to the nuanced nature of language and potential biases in training data; however, semi-automated systems can significantly support human moderators.

In this chapter, we explore both prior work and the current state of the art in detecting hate speech, misogyny, and sexism across various modalities, including text and multimedia content. We discuss methodologies and technologies that have been employed, highlighting both their strengths and limitations. Moreover, we propose a novel approach to improving the detection of misogynistic and sexist content on TikTok, a platform known for its dynamic and diverse multimedia content. TikTok videos, which often combine text, audio, and visual elements, present unique challenges for automated detection systems. Our approach leverages the integration of multimodal information and recent advances in Natural Language Processing (NLP) and Computer Vision to address these challenges.

2.2 Textual Work

The field of NLP has increasingly focused on detecting hate speech and sexism, driven by their growing societal impacts, especially on social platforms. One foundational effort provided a corpus of misogynous tweets, labeled from various perspectives, and explored NLP features and machine learning models for detecting and classifying misogynistic language [3]. Building upon these early approaches, SemEval-2019 Task 5 targeted hate speech against immigrants and women [6]. This included a binary classification task (Subtask A) to identify the presence of hate speech and a finer-grained classification task (Subtask B) to detect features within hateful content, such as aggression and whether the target is an individual or a group. The challenge saw 108 submissions for Subtask A and 70 for Subtask B from 74 teams. Further extending these efforts, the AMI challenge at Evalita2020 evaluated misogyny and aggressiveness in Italian tweets, receiving a total of 20 runs for Subtask A and 11 for Subtask B, submitted by 8 teams [14]. More recent initiatives, like SemEval-2023 Task 10, developed a hierarchical taxonomy of sexist content and curated a dataset of 20,000 social media comments to enhance the explainability of detection

[1] `https://www.oed.com/view/Entry/misogyny`
[2] `https://www.oed.com/view/Entry/sexism`

methods [17]. This progression underscores a continuous commitment to refining methods for tackling hate speech and sexism on digital platforms.

Recent contributions in this area include [19], which evaluated biases in abusive language detection systems using BERT-based models, highlighting performance issues related to fairness and bias. The evaluation shows that, although BERT-based classifiers achieve high accuracy levels on a variety of natural language processing tasks, they perform very poorly regarding fairness and bias, particularly on samples involving implicit stereotypes, expressions of hate towards minorities, and protected attributes such as race or sexual orientation. In [1], new state-of-the-art results in hate speech detection were achieved using T5 models, data augmentation, and ensemble techniques. They achieved new SoTA on two subtasks - macro F1 scores of 91.73% and 53.21% for subtasks A and B of the HASOC 2020 dataset[3], where previous SoTA were 51.52% and 26.52%, respectively. In [10], a multi-target approach for hate speech detection was developed using sentic computing resources and neural models. The study focuses on: (1) transferring knowledge from general to specific instances of hate speech; (2) detecting hate speech across various topics and targets with greater detail; and (3) using affective knowledge from resources like SenticNet[4] and HurtLex[5] to identify specific manifestations of hate speech.

In [8], a topic-oriented method to enhance generalization in hate speech detection was proposed, which improved model reliability across various datasets. The authors introduced an innovative yet straightforward approach to more precisely identify which topics are most effectively captured in implicit expressions of hate. They demonstrated that choosing combinations of datasets with broader out-of-domain topical coverage enhances the reliability of automatic hate speech detection. Furthermore, in [29], the Measuring Hate Speech corpus was introduced, along with the development of a high-accuracy machine learning model for detecting hate speech on Twitter. The authors implemented machine learning algorithms such as Logistic Regression, achieving 92% accuracy. They also created a Support Vector Machine (SVM) algorithm and employed a Topic Modeling technique to identify topics within the corpus.

In [30], BERT-based models for explainable online sexism detection were developed, achieving high F1 scores in the SemEval-2023 Task 10 competition. In [24], the effectiveness of conventional machine learning techniques in detecting online sexist content was explored, highlighting strengths and weaknesses of different classifiers. Recent studies have made significant advances in this field, for instance, in [33], the first dataset of sexist expressions on Twitter in Spanish was created, demonstrating the feasibility of using machine learning for sexism detection. The results show that sexism is frequently found in many forms in social networks, including a wide range of behaviors, and can be detected using deep learning approaches. In [23], a multilingual and cross-domain study on misogyny detection on Twitter was conducted, outperforming state-of-the-art systems in benchmark datasets. The authors investigated the most important features to detect misogyny and the issues contributing to the difficulty of misogyny detection by proposing a novel system and conducting a broad evaluation on this task. They also studied the relationship between misogyny and other abusive language phenomena through cross-domain classification experiments and explored the feasibility of detecting misogyny in a multilingual environment via cross-lingual classification experiments. Lastly, in [38], an ensemble of 18 models, including DeBERTa and BERT variations, was leveraged for identifying sexist text with high F1 scores in the SemEval-2023 Task 10 competition. Since 2021, the EXIST (sEXism Identification in Social neTworks)[6] task addresses the problem of sexism identification in social networks, capturing sexism in a broad sense. This serie of scientific events and shared tasks aims to identify sexism ranging from explicit misogyny to subtle expressions involving implicit sexist behaviors [31, 32, 26]. 50 teams from over 15 countries submitted their results, achieving notable success, particularly in the sexism detection task. Despite these achievements, there remains a need for improvement, especially in categorizing sexism based on which aspect of women is being targeted and undermined.

2.3 Multimodal Work

Recent advancements in multimodal analysis have significantly enhanced the detection of hate speech in memes and images. The Multimodal Hate Speech Event Detection task organized in 2023 explored binary and target-specific detection strategies in text-embedded images, demonstrating the effectiveness of multimodal approaches in identifying hate speech [36]. The shared task had two subtasks. Sub-task A required participants to pose hate speech detection as a binary problem, i.e., they had to detect whether

[3] https://hasocfire.github.io/hasoc/2020/dataset.html
[4] https://sentic.net/
[5] https://github.com/valeriobasile/hurtlex
[6] http://nlp.uned.es/exist2024/

the given text-embedded image had hate or not. Sub-task B required participants to identify the targets of the hate speech, namely individual, community, and organization targets in text-embedded images. For both subtasks, the participants were ranked based on the F1 score. The best F1 scores in sub-task A and sub-task B were 85.65 and 76.34, respectively. A novel method introduced in 2023 utilizes pre-trained vision-language models (PVLMs) for hateful meme detection [9]. The authors propose a probing-based captioning approach to leverage PVLMs in a zero-shot visual question answering (VQA) manner. Specifically, they prompt a frozen PVLM by asking hateful content-related questions and use the answers as image captions.

Additionally, a study conducted in 2018 demonstrated the superiority of a multimodal approach over unimodal methods in detecting sexist content in advertisements [15]. It proved that a multimodal approach that considers the trained visual classifier and a textual one permits good classification performance on the second dataset, reaching 87% recall and 75% accuracy, which are significantly higher than the performance obtained by each of the corresponding unimodal approaches. The Multimedia Automatic Misogyny Identification (MAMI) task at SemEval-2022 focused on identifying misogynous content in memes [13]. The task was organized into two related subtasks: the first focused on recognizing whether a meme is misogynous or not (Sub-task A), while the second was devoted to recognizing types of misogyny (Sub-task B). MAMI was one of the most popular tasks at SemEval-2022 with more than 400 participants, 65 teams involved in Sub-task A and 41 in Sub-task B from 13 countries. Further studies in multimodal analysis include those in [4], which reported a 71% F-score in identifying misogynous memes using a multimodal system based on the CLIP model. In [34], transformer models were used to detect misogynous content in text and images, achieving better results with an ensemble of ALBERT and CNN models. In [25], the UNITER model was enhanced with image sentiment and graph convolutional networks for multimedia automatic misogyny identification.

In recent years, significant advances have been made in the detection of hate speech in videos. In [35], a combined approach using images, audio, and textual features for hate speech detection in videos was proposed, achieving significant improvements in detection accuracy. In [12], a multimodal approach combining acoustic and textual elements was applied to detect Amharic hate speech, achieving an accuracy of 88.15% with BILSTM. In [28], a multimodal deep learning framework combining auditory and semantic features was proposed to detect hateful multimedia content, showing significant improvement over text-based models. In [20], the use of facial expressions in hate speech detection was investigated, achieving a validation accuracy of 84.8% with a Random Forest classifier. In [37], a method that classifies videos into normal or hateful categories by exclusively analyzing the transcriptions of spoken content was developed. Another important study published a dataset of Portuguese videos, which primarily includes textual features extracted from the video content [2]. Furthermore, in [11], a dataset was created by manually annotating around 43 hours of BitChute[7] videos. Their multimodal approach led to a 5.7% improvement in F1-score over systems that used only one modality. To our knowledge, no research has been conducted on the detection of sexism in videos.

2.4 Proposal

The primary goal of our work is to advance the capabilities of sexism detection by extending the use of multimodality beyond traditional text and image analysis to include video content, with a specific focus on videos from platforms like TikTok. While considerable research has been conducted on identifying sexism in textual content and images, the domain of video content, particularly that which circulates on social media platforms, has not been thoroughly explored.

Our approach builds on the insights gained from prior studies that have demonstrated the effectiveness of leveraging multiple modalities, such as text and images, to detect sexism. By incorporating these modalities, we aim to develop a more nuanced and effective method for identifying sexist content in videos. To facilitate this, we have compiled the first dataset of TikTok videos specifically designed for analyzing sexism in both English and Spanish. This dataset not only serves as a foundational resource for our current analysis but also establishes a benchmark for future research and for comparing different models.

In addition to integrating traditional modalities like text and images, our study uniquely incorporates linguistic features, emotional cues, audio, and video components to identify indicators of sexist content. By analyzing how these different elements interact and contribute to the overall context of sexism within

[7] `https://www.bitchute.com/`

videos, we aim to uncover subtle and overt expressions of sexism that may be overlooked by unimodal detection approaches.

To validate the efficacy of our multimodal approach, we compare it against unimodal models that rely on single types of input. This comparison helps us assess the added value of integrating multiple data streams in enhancing the detection accuracy and understanding of sexist content. By employing and refining these multimodal strategies, we seek to contribute significantly to the field, providing robust tools that can assist in moderating and reducing sexist content on widely used social media platforms like TikTok.

2.5 Legal and Ethical Considerations

The extraction and analysis of TikTok videos for detecting sexism raise significant legal and ethical concerns. It is crucial to navigate these considerations carefully to ensure compliance with relevant laws and respect for users' rights.

One of the primary legal concerns is the privacy of TikTok users. Videos posted on social media platforms often contain personal information, either explicitly or implicitly. Therefore, it is essential to ensure that the data used for analysis is obtained in a manner that respects users' privacy rights. This includes anonymizing any personal identifiers and obtaining proper consent when necessary. Additionally, researchers must comply with the platform's terms of service and privacy policies.

The development and deployment of automated systems to detect sexist content must be guided by ethical principles to avoid unintended consequences. These systems should be designed to minimize biases and ensure fairness. Bias in training data can lead to discriminatory outcomes, particularly against marginalized groups. Therefore, it is essential to use diverse and representative datasets and continually evaluate and update models to mitigate bias. Furthermore, automated systems should be transparent and accountable. Users should be informed about how their data is used and the criteria for flagging content. Mechanisms for appeal and review should be in place to address any potential errors or unjust flagging of content.

While it is crucial to combat sexism on social media, it is equally important to balance this goal with the protection of freedom of expression. Automated detection systems must be carefully calibrated to avoid over-censorship, ensuring that legitimate content is not inadvertently removed or flagged. Striking this balance requires ongoing evaluation and refinement of detection algorithms to ensure they are both effective and fair.

Finally, the ethical use of data extends to how research findings are applied. The goal should be to create a safer and more inclusive online environment without infringing on users' rights or freedoms. Researchers and developers must remain mindful of the potential impacts of their work and strive to use their findings responsibly to inform policy and improve moderation practices.

Chapter 3

Dataset, Servipoli Annotation and Tasks

3.1 Video Extraction Using Apify's TikTok Hashtag Scraper

The process of creating the TikTok dataset commenced with the collection of videos from the platform. To facilitate this, the Tiktok Hashtag Scraper tool, provided by Apify, a platform specializing in web scraping and data extraction, was utilized[1]. Figure 3.1 showcases examples of the TikToks collected, with the top being in Spanish and the bottom in English.

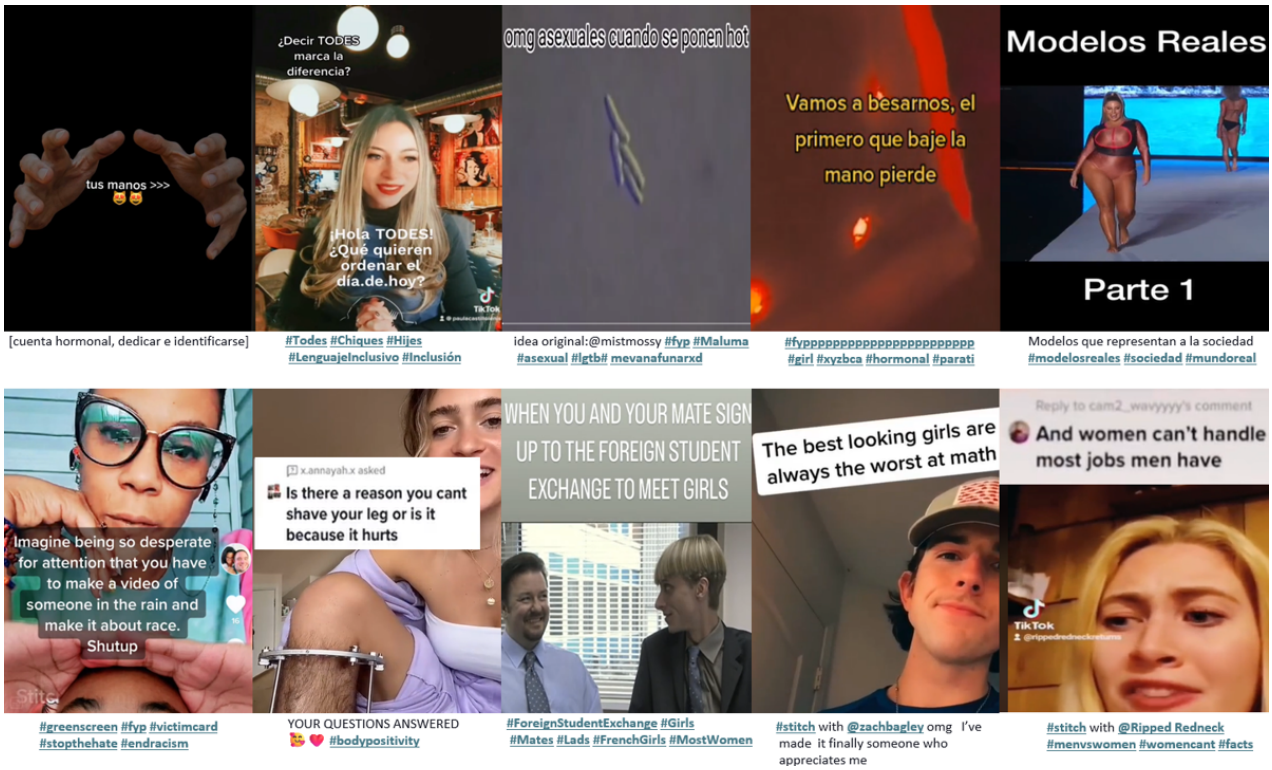


Figure 3.1: Examples of TikToks collected: top in Spanish and bottom in English.

[1] https://apify.com/clockworks/tiktok-hashtag-scraper

A total of 185 Spanish hashtags and 61 English hashtags were selected, all deemed relevant and potentially associated with sexist content. Examples of the Spanish hashtags include *#envras*, *#laspibas*, *#nosoyfeminista*, *#pibasvspibes* and *#nosequienlasinvento*. The English hashtags selected include *#girlsvsboys*, *#golddigger*, *#notallwomen* and *#menvswomen*. Figure 3.2 presents the WordClouds of these hashtags in both Spanish and English.

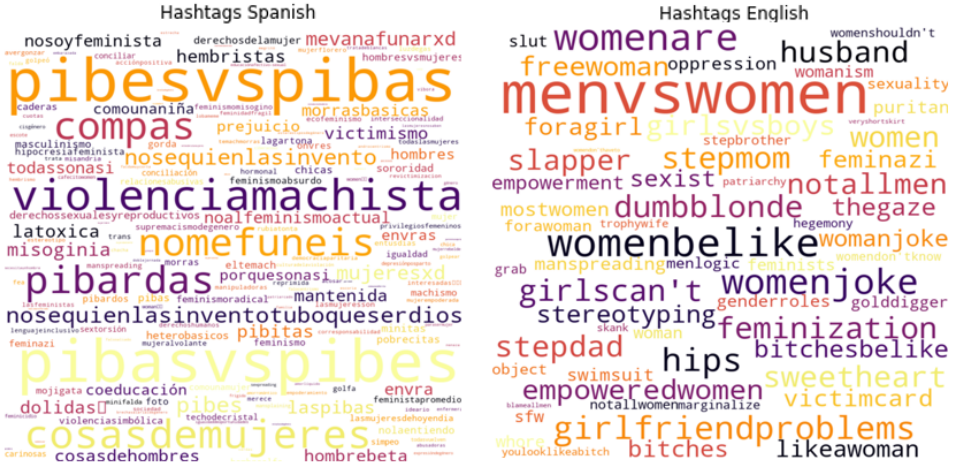


Figure 3.2: WordClouds of Hashtags in Spanish and English

3.2 Tasks and Servipoli Annotation Process

3.2.1 Tasks for Detecting and Categorizing Sexism in TikTok Videos

In line with the objectives of the EXIST initiative, which aims to identify sexism in social networks, this research focuses on detecting and categorizing sexism in TikTok videos. This involves recognizing a broad spectrum of sexist expressions, ranging from explicit misogyny to subtler forms of gender bias. This study is structured into three tasks aimed at analyzing and classifying various aspects of sexism present in TikTok videos:

1. **Sexism Detection.**
   This task involves examining TikTok videos to determine whether they contain sexist content. It's a binary classification task where systems must make a decision about whether a video is sexist or not.

   - **Sexist.** A TikTok video that discusses, portrays, or addresses gender-related stereotypes, roles, or issues.
   - **Not Sexist.** A TikTok video that does not focus on gender-related stereotypes, roles, or issues.

2. **Source Intention Classification.**
   In this task, TikTok videos are categorized based on the intention behind their creation. The categories include direct sexist messages and reported messages (combining reported and judgmental messages).

   - **Direct Sexist.** A TikTok video explicitly promoting gender stereotypes or perpetuating sexist beliefs, such as suggesting women belong in certain roles or positions.
   - **Reported Sexist.** A TikTok video sharing personal experiences of encountering sexism or misogyny, either directly witnessed or reported by others.

3. **Sexism Categorization.**
   This task involves classifying each sexist TikTok video based on the aspect of women's lives targeted by the sexism. Categories include ideological and inequality, role stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence.

- **Ideological and Inequality.** A TikTok video discrediting feminist movements or reinforcing societal inequalities by belittling women's rights or contributions.
- **Role Stereotyping and Dominance.** A TikTok video perpetuating stereotypes about gender roles, suggesting that women are better suited for certain tasks or professions.
- **Objectification.** A TikTok video portraying women solely as objects of desire, focusing on their physical appearance rather than their personality or abilities.
- **Sexual Violence.** A TikTok video containing explicit content or harassment of a sexual nature, including comments or behaviors that promote or condone sexual assault.
- **Misogyny and Non-sexual Violence.** A TikTok video expressing hatred or hostility towards women, including verbal abuse, threats, or acts of violence.

3.2.2 Servipoli Annotation Process

The annotation was conducted using Servipoli's service[2]. A total of eight students were selected for this task, organized in pairs. Each pair of students was assigned to annotate 1000 TikToks, either in Spanish or English, as appropriate.

Before commencing with the annotation, all eight students were provided with a detailed explanation of the three tasks involved in the study. During this briefing session, all doubts and questions related to the labeling criteria were clarified to ensure consistent understanding among all annotators.

To validate the accuracy and consistency of the annotations, a preliminary test was conducted using a set of 10 TikToks. This test served as a training exercise to familiarize the students with the annotation process and to ensure that they were able to correctly apply the labeling criteria for each task.

In cases where there was disagreement between the annotations provided by the pairs of students who labeled the same set of 1000 TikToks, the final decision was made by a member of our research team. Disagreements were primarily considered in the first two binary tasks: determining whether the video was sexist and agreeing on the intention behind the video, in cases where both annotators had labeled the video as sexist but differed in intention. Regarding the categorization of the aspects of sexism, the labels provided by both annotators often did not align fully across all specified categories due to the complexity and diversity of sexism. To address this, labels were merged to capture a broader spectrum of expressions, or a decision was made by a member of our team in cases of disagreement regarding sexism or intention.

Annotators	Sexist %	Intention %	Not Found %
1 vs 5 (EN1)	25.93	6.75	6.57
2 vs 6 (EN2)	21.52	4.94	6.17
3 vs 7 (ES1)	18.92	3.21	3.04
4 vs 8 (ES2)	25.00	1.79	5.10

Table 3.1: Estimated disagreements between pairs of annotators for English (EN) and Spanish (ES).

Table 3.1 presents the estimated disagreements between pairs of annotators in three key aspects: labeling TikToks as sexist, identifying the intention behind the videos, and encountering where one annotator could not view the content properly.

The highest disagreements were observed in determining whether a TikTok video was sexist or not. For instance, in pair 1 vs 5 (EN1), 25.93% of TikToks had conflicting annotations regarding their sexist nature. In contrast, there seemed to be more agreement among annotators when identifying the intention behind the videos, with generally lower percentages of disagreement across all pairs, such as 6.75% in pair 1 vs 5 (EN1).

The last column, "Not Found %," refers to the percentage of TikToks in which one annotator was unable to view the content correctly. This could be due to reasons such as missing audio, deleted videos, private account settings, or content in a language other than the one they were assigned to annotate. These issues were particularly notable in pair 1 vs 5 (ES2), where 6.57% of TikToks were not accessible to one annotator.

This rigorous annotation process, involving multiple annotators and validation steps, was implemented to ensure that the labels assigned to each TikTok were consistent, reliable, and properly reflected the various facets of sexism identified in the study.

[2] `https://www.servipoli.es/`

3.2.3 Cohen's Kappa Agreement Scores Across Different Tasks

Cohen's Kappa was utilized to evaluate Inter-Annotator Agreement (IAA) across the two tasks. In the *sexism detection* task, the average Kappa value between annotator pairs was 0.499, indicating moderate agreement ($\kappa \in [0.41, 0.60]$). The *source intention classification* task showed substantial agreement with an average Kappa of 0.672 ($\kappa \in [0.61, 0.80]$).

While the values of IAA can be considered good in the two first tasks, the are low in the third task that involves the categorization into five categories of sexism. Here, the average Kappa values across categories were: Ideological Inequality ($\kappa = 0.306$), Stereotyping-Dominance ($\kappa = 0.409$), Objectification ($\kappa = 0.312$), Sexual Violence ($\kappa = 0.396$), and Misogyny-Non-Sexual Violence ($\kappa = 0.179$).

Although the dataset employed in the experiments was entirely annotated by human annotators, we were interested in investigating the Kappa values for annotations made by GPT-3.5 Turbo. The average Kappa values for the *sexism detection* task and the *source intention classification* task by GPT-3.5 Turbo were 0.282 and 0.246, respectively. These results, which show fair agreement ($\kappa \in [0.21, 0.40]$), indicate that while GPT-3.5 Turbo achieves a basic level of concordance with human annotations, it does not yet reach the agreement levels of human annotators. This is why we did not consider including GPT-3.5 Turbo as an additional annotator. Figures 3.3 and 3.4 illustrate the Cohen's Kappa agreement scores achieved by GPT-3.5 Turbo.

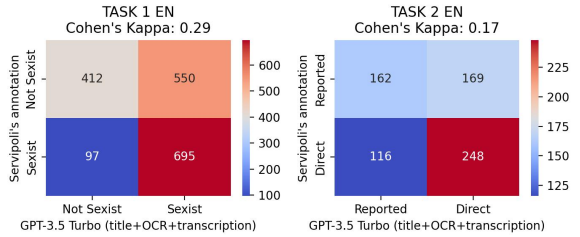


Figure 3.3: Cohen's Kappa agreement for English TikToks

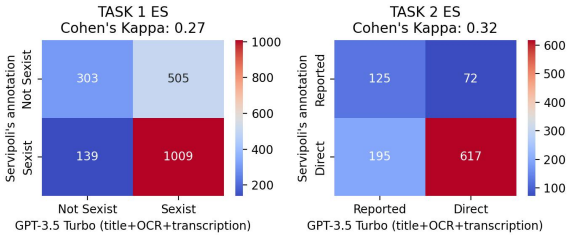


Figure 3.4: Cohen's Kappa agreement for Spanish TikToks

3.3 Summary of TikTok Corpus

3.3.1 Spanish TikTok Corpus

The Spanish TikTok corpus consists of a total of 1969 TikToks, with a cumulative duration of 13.863 hours. Among these, 817 (41.49%) are categorized as non-sexist, while 1152 (58.51%) are considered sexist. There is a significant imbalance between the reported sexist content, which accounts for 362 (31.42%), and directed sexist content, which comprises 790 (68.58%).

The average duration of the TikToks in this corpus is 25.35 seconds. Notably, the reported sexist content tends to have a longer average duration, with a mean of 34.07 seconds.

When examining the distribution of the videos across different categories, the Stereotyping-Dominance category is predominant, making up 62.67% of the sexist videos. The least represented category is Sexual Violence, especially within the directed sexism content, where only 54 videos (6.84%) are categorized as such. For detailed statistics, refer to Table 3.2.

3.3.2 English TikTok Corpus

The English TikTok corpus comprises a total of 1773 TikToks, with a cumulative duration of 11.827 hours. Out of these, 975 (55%) are non-sexist, while 798 (45%) are categorized as sexist. Similar to the Spanish corpus, there is an imbalance between the reported sexist content, which constitutes 297 (37.22%), and directed sexist content, which accounts for 501 (62.78%).

The average duration of the TikToks in this corpus is 24.02 seconds. As observed in the Spanish corpus, the reported sexist content tends to have a longer average duration, with a mean of 29.89 seconds.

The Stereotyping-Dominance category is predominant in this corpus as well, accounting for 80.58% of the sexist videos. The category with the least representation is Sexual Violence, especially within the directed sexism content, where only 18 videos (3.59%) are categorized as such. For detailed statistics, refer to Table 3.3.

Table 3.2: Statistics of the Spanish TikTok Corpus

	Non-sexist	**Reported**	**Direct**	**Total**
Count (%)	817 (41.49%)	362 (31.42%)	790 (68.58%)	1969
Total Duration (hours)	5.095	3.426	5.343	13.863
Mean Duration ± Std (seconds)	22.45 ± 16.21	34.07 ± 19.71	24.35 ± 16.83	25.35 ± 17.65
Count (%) by Category				
Ideological Inequality	-	140 (38.67%)	204 (25.82%)	344 (29.86%)
Stereotyping-Dominance	-	186 (51.38%)	536 (67.84%)	722 (62.67%)
Objectification	-	47 (12.98%)	156 (19.75%)	203 (17.62%)
Sexual Violence	-	95 (26.24%)	54 (6.84%)	149 (12.93%)
Misogyny-Non-Sexual Violence	-	89 (24.59%)	134 (16.96%)	223 (19.36%)
Total Duration by Category (hours)				
Ideological Inequality	-	1.426	1.514	2.940
Stereotyping-Dominance	-	1.772	3.776	5.549
Objectification	-	0.358	0.939	1.297
Sexual Violence	-	0.944	0.324	1.267
Misogyny-Non-Sexual Violence	-	0.899	0.756	1.655
Mean Duration by Category ± Std (seconds)				
Ideological Inequality	-	36.67 ± 18.42	26.72 ± 18.60	30.77 ± 19.13
Stereotyping-Dominance	-	34.31 ± 19.75	25.36 ± 17.06	27.67 ± 18.21
Objectification	-	27.39 ± 19.01	21.67 ± 16.02	23.00 ± 16.88
Sexual Violence	-	35.75 ± 19.69	21.58 ± 14.19	30.62 ± 19.11
Misogyny-Non-Sexual Violence	-	36.36 ± 19.32	20.31 ± 15.04	26.72 ± 18.59

Table 3.3: Statistics of the English TikTok Corpus

	Non-sexist	**Reported**	**Direct**	**Total**
Count (%)	975 (54.99%)	297 (37.22%)	501 (62,78%)	1773
Total Duration (hours)	5.880	2.466	3.482	11.827
Mean Duration ± Std (seconds)	21.71 ± 16.23	29.89 ± 20.49	25.02 ± 18.47	24.02 ± 17.89
Count (%) by Category				
Ideological Inequality	-	180 (60.61%)	158 (31.54%)	338 (42.36%)
Stereotyping-Dominance	-	252 (84.85%)	391 (78.04%)	643 (80.58%)
Objectification	-	91 (30.64%)	135 (19.75%)	226 (28.32%)
Sexual Violence	-	53 (17.85%)	18 (3.59%)	71 (8.90%)
Misogyny-Non-Sexual Violence	-	46 (15.49%)	57 (11.38%)	103 (12.91%)
Total Duration by Category (hours)				
Ideological Inequality	-	1.529	1.282	2.810
Stereotyping-Dominance	-	2.001	2.839	4.840
Objectification	-	0.947	0.859	1.806
Sexual Violence	-	0.543	0.099	0.642
Misogyny-Non-Sexual Violence	-	0.492	0.395	0.886
Mean Duration by Category ± Std (seconds)				
Ideological Inequality	-	30.57 ± 29.21	18.81 ± 18.60	29.93 ± 19.93
Stereotyping-Dominance	-	28.59 ± 26.14	18.87 ± 17.06	27.10 ± 19.03
Objectification	-	37.46 ± 19.59	22.92 ± 16.92	28.77 ± 19.37
Sexual Violence	-	36.89 ± 20.81	19.79 ± 16.41	32.55 ± 21.05
Misogyny-Non-Sexual Violence	-	38.49 ± 20.38	24.92 ± 19.51	30.98 ± 20.93

Chapter 4

Text, Audio, Video, and Multimodal Models

4.1 Text Models

4.1.1 Data Preprocessing

In the preprocessing phase of TikTok text data, several steps have been taken into account to ensure an accurate and unbiased representation of the content:

- **Hashtags and User Mentions.** Hashtags (#) and user mentions (@) have been removed from the titles. These tags can introduce significant biases when determining if a video is sexist or when inferring its intent or category.
- **Emojis.** Emojis in the titles have been retained. To interpret the natural language meaning of any emoji, the *emoji* library has been employed[1].
- **Transcription.** Obtaining the video transcription is crucial for a detailed analysis of the verbal content. For this purpose, the *clu-ling/whisper-large-v2-spanish*[2] model has been used for Spanish TikToks, and *openai/whisper-large-v3*[3] for English TikToks. Whisper is a pre-trained model for automatic speech recognition (ASR) and speech translation. Trained on 680k hours of labelled data, Whisper models demonstrate a strong ability to generalize to many datasets and domains without the need for fine-tuning [27].
- **Optical Character Recognition (OCR).** Detecting all possible OCR from the videos is another essential part of the preprocessing. For this purpose, the *easyocr* library[4] in Python has been chosen, which supports OCR with 80+ languages and all popular writing scripts. The idea is to extract the text from each TikTok every 30 frames and then perform post-processing with GPT-3 to clean possible repetitions or grammatical errors.

Although this preprocessing has yielded good results, it is not perfect. There may be noise in the transcription or OCR, or certain information that we might be missing. An example of the title, OCR, and transcription of a TikTok is shown in Figure 4.1.

[1] `https://pypi.org/project/emoji/`
[2] `https://huggingface.co/clu-ling/whisper-large-v2-spanish`
[3] `https://huggingface.co/openai/whisper-large-v3`
[4] `https://pypi.org/project/easyocr/`

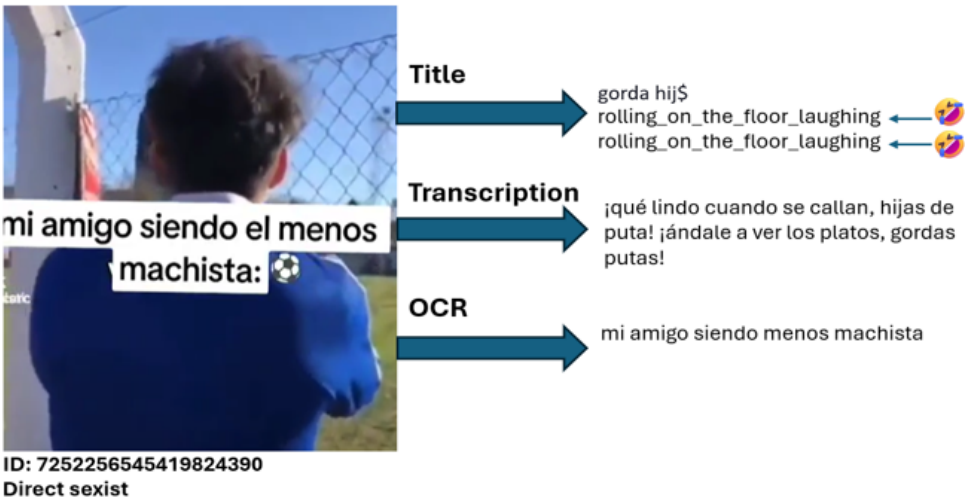


Figure 4.1: Example of the title, OCR, and transcription of one TikTok

4.1.2 Linguistic Resources

In this section, various linguistic resources employed for text analysis in the study are introduced. These encompass LIWC[5], HURTLEX, and the use of pre-trained Transformers to extract emotions or variables associated with hate speech. Features are derived by amalgamating the title, transcription, and OCR of the TikToks, which will be utilized in subsequent models.

As part of the exploratory analysis, statistical tests using the Mann-Whitney U test [21] are executed. This non-parametric test, which does not presuppose normality, is applied to discern significant differences between the means of various features for sexist versus non-sexist videos, as well as reported versus direct sexism. It is worth noting that these statistical tests are exclusively conducted for the TikToks in Spanish to offer a more targeted analysis.

4.1.2.1 LIWC

Linguistic Inquiry and Word Count (LIWC)[18] is a text analysis tool that categorizes words into specific psychological and social categories. It's used to understand and analyze language patterns related to emotions, social relationships, and cognitive styles. LIWC contains over 100 dictionaries, each with words and phrases associated with different psychological dimensions, helping researchers link language to human behavior and emotions.

For Spanish text analysis, we will use the 2007 version of the LIWC dictionary. This version is tailored specifically for the Spanish language, enabling us to categorize words based on their psychological and social meanings in Spanish. For English and texts translated into English, we will employ the 2015 version of the LIWC dictionary. For more details on the Spanish version, please refer to Figure A.2.

Two tables are presented to highlight the significant differences in LIWC features between various categories of Spanish TikToks. In Table 4.1, significant differences were observed between sexist and non-sexist TikToks. Sexist TikToks displayed a higher frequency of sexual terms "Sexual" (0.85 vs 0.49), affective terms "Afect" (4.92 vs 4.42), and social terms "Social" (11.70 vs 10.76). On the other hand, non-sexist TikToks exhibited a greater use of achievement-related terms "Logro" (1.96 vs 1.55) and terms that indicate position in relation to another point "Relativ" (8.35 vs 7.59).

In Table 4.2, significant differences were found between TikToks with reported and direct sexism. TikToks with reported sexism had a more prominent use of prepositions (11.62 vs 9.97), inclusive terms "Incl" (5.75 vs 5.09), and a higher frequency of the verb "Nos" (0.42 vs 0.31). Conversely, TikToks with direct sexism exhibited a more frequent use of the pronoun "Yo" (2.39 vs 1.78), terms related to "Amigos" (0.44 vs 0.26), and a more significant use of profane words "Maldec" (0.40 vs 0.22).

In summary, these tables and graphs provide valuable insights into the linguistic patterns distinguishing non-sexist TikToks from sexist ones and reported sexism from direct sexism.

[5] https://www.liwc.app/

Table 4.1: Significant Differences in LIWC Features for Sexist vs Non-Sexist Spanish TikToks. Bold values indicate the group with the highest mean for each feature.

Feature	p-value	Non-sexist Mean	Non-sexist SD	Sexist Mean	Sexist SD
TotPron	0.0145	**16.13**	6.02	15.28	6.19
PronPer	0.0036	**10.59**	5.14	9.80	5.48
Articulo	$7.53 \cdot 10^{-6}$	8.18	4.86	**9.51**	5.19
MecCog	0.0318	17.69	7.04	**18.66**	6.87
Ellos	0.0471	1.77	2.03	**1.90**	1.96
Social	0.0006	10.76	5.73	**11.70**	5.13
Afect	0.0374	4.42	3.02	**4.92**	3.49
EmoNeg	$8.28 \cdot 10^{-5}$	1.71	1.99	**2.32**	2.78
Enfado	$3.93 \cdot 10^{-9}$	0.75	1.44	**1.25**	1.70
Logro	0.0109	**1.96**	2.15	1.55	1.95
Relativ	0.0126	**8.35**	4.41	7.59	3.81
Maldec	0.0091	0.20	0.64	**0.33**	0.86
Discrep	0.0008	1.48	1.67	**1.84**	1.88
Insight	0.0470	2.27	1.97	**2.54**	2.25
Humanos	$4.10 \cdot 10^{-17}$	1.14	1.52	**2.31**	2.47
Negacio	0.0022	2.83	5.24	**3.11**	4.25
verbELLOS	$2.10 \cdot 10^{-5}$	0.59	1.19	**0.83**	1.26
Yo	0.0009	**2.66**	2.84	2.16	2.83
verbYO	0.0025	**1.60**	1.96	1.21	1.53
Oir	0.0023	1.03	1.52	**1.21**	1.50
Sexual	$1.73 \cdot 10^{-6}$	0.49	1.00	**0.85**	1.81
VosUtds	0.0144	0.12	0.39	**0.21**	0.66
Muerte	0.0022	0.15	0.47	**0.27**	0.72
Triste	0.0378	**0.23**	0.69	0.22	0.53

Table 4.2: Significant Differences in LIWC Features between Reported and Direct Sexism Spanish TikToks. Bold values indicate the group with the highest mean for each feature.

Feature	p-value	Reported Mean	Reported SD	Direct Mean	Direct SD
Prepos	1.88×10^{-7}	**11.62**	4.39	9.97	4.77
Incl	0.00219	**5.75**	2.71	5.09	2.83
EmoNeg	0.00306	**2.52**	2.25	2.20	3.07
Enfado	4.57×10^{-5}	**1.49**	1.70	1.09	1.68
Espacio	0.0143	**3.39**	2.23	3.09	2.47
Maldec	0.0365	0.22	0.65	**0.40**	0.97
VerbAux	0.00180	**0.97**	1.57	0.65	1.04
Pasado	0.00970	**1.07**	1.32	1.01	1.62
EmoPos	0.0477	2.35	1.92	**2.98**	3.05
Amigos	0.0381	0.26	0.59	**0.44**	0.90
Subjuntiv	0.00696	**1.93**	1.94	1.54	1.70
Trabajo	0.0136	**1.84**	2.05	1.67	2.24
Yo	0.00216	1.78	2.59	**2.39**	2.95
Sexual	0.00227	**0.92**	1.29	0.80	2.07
Dinero	0.00559	**0.52**	0.88	0.45	0.97
Inhib	0.000169	**0.62**	0.89	0.48	0.88
verbNOS	0.000350	**0.42**	0.82	0.31	0.84
Nosotro	6.96×10^{-5}	**0.51**	1.12	0.29	0.85
Futuro	0.00455	**0.0119**	0.102	0.0	0.0

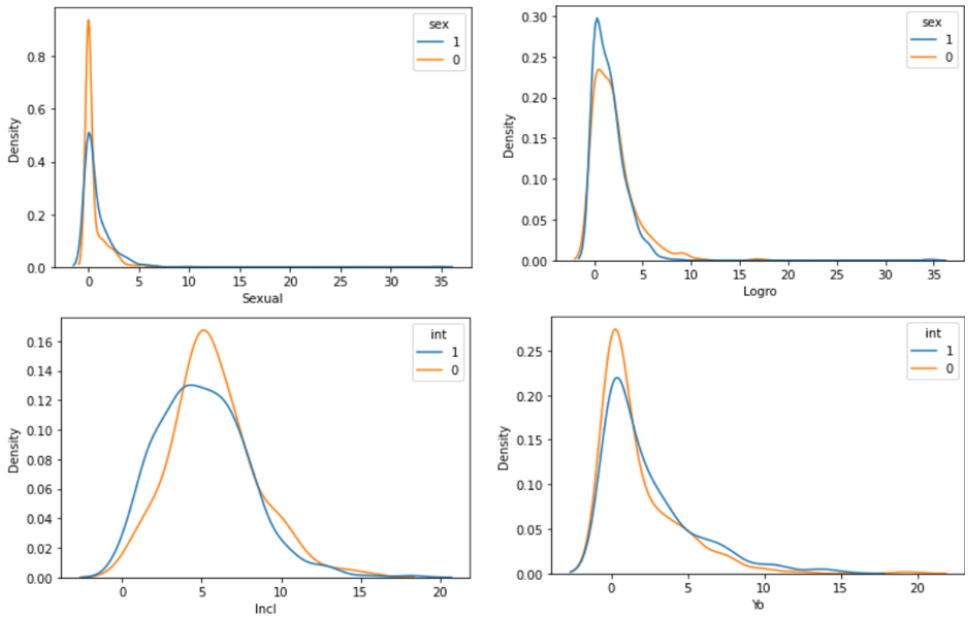


Figure 4.2: Distributions of some features of LIWC

The top graphs of 4.2 illustrate the frequency distributions of terms related to sexual content and achievement ("Sexual" and "Logro") across non-sexist and sexist TikToks. The bottom graphs showcase the frequency distributions of terms related to inclusivity and self-reference ("Incl" and the use of "Yo") for direct vs reported sexist TikToks.

4.1.2.2 HURTLEX

HurtLex is a multilingual lexicon that encompasses offensive, aggressive, and hateful words in over 50 languages. The lexicon classifies these terms into 17 distinct categories, complemented by a macro-category to denote the presence of stereotypes. Structurally, HurtLex operates on a 2-level system: **Conservative**, where translations are based on the offensive senses of the original lexicon words, and **Inclusive**, where translations encompass all potentially relevant senses from the original lexicon. Specific examples of these categorized words in Spanish are shown in Table 4.3. A comprehensive description of HurtLex can be found in [7].

Table 4.3: HurtLex Spanish Table

Category	Count	Explanation	Example
CDS	1285	Derogatory Words	ladrón
AN	679	Words Related to Animals	bison
DMC	361	Moral and Behavioural Defects	maldad
QAS	349	Words with Potential Negative Connotations	cojon
DDP	332	Phsyical Disabilities and Diversity	idiotez
ASM	328	Words Related to Male Genitalia	tacto suave
SVP	322	Words related to the Seven Deadly Sins	aburrido
RE	272	Felonies and Words Related to Crime	villano
OM	213	Words Related to Homosexuality	posaderas
PS	203	Negative and Stereotypes and Ethnic Slurs	beduin
OR	173	Words Related to Plants	naranja quemado
PR	165	Words Related to Prostitution	añojo
PA	109	Profession and Occupation	controlador
ASF	90	Words Related to Female Genitalia	panoli
IS	75	Words Related to Social and Economic Disadvantage	parsimonia
DDF	36	Cognitive Disabilities and Diversity	minusvalido
RCI	14	Location and Demonyms	bárbaro

Looking at the differences between sexist and non-sexist Spanish TikToks (Table 4.4), the videos categorized as sexist show a higher mean percentage across almost all categories compared to non-sexist videos. Notably, the categories with the most pronounced differences include 'an' (Words Related to Animals), 'pr' (Words Related to Prostitution), 're' (Felonies and Words Related to Crime), 'asf' (Words Related to Female Genitalia), and 'qas' (Words with Potential Negative Connotations). This indicates that sexist TikToks tend to incorporate derogatory terms related to animals, prostitution, crime, female genitalia, and potential negative connotations more frequently than non-sexist TikToks. Interestingly, the category 'dmc' (Moral and Behavioural Defects) shows a significant difference in the opposite direction, being more pronounced in non-sexist TikToks. This suggests that non-sexist TikToks might emphasize moral and behavioral defects more frequently than their sexist counterparts.

On the other hand, the differences between reported and direct sexism in Spanish TikToks (Table 4.5) reveal contrasting patterns. Reported sexism shows higher mean percentages in 'is' (Words Related to Social and Economic Disadvantage), 're' (Felonies and Words Related to Crime), and 'ddp' (Physical Disabilities and Diversity). This suggests that when sexism is reported, it tends to be associated more with social and economic disadvantages, crimes, and physical disabilities. In contrast, direct sexism, exhibits higher mean percentages in 'cds' (Derogatory Words), 'pr' (Words Related to Prostitution), 'dmc' (Moral and Behavioural Defects), and 'svp' (Words related to the Seven Deadly Sins). This implies that direct expressions of sexism are more likely to include derogatory terms, references to prostitution, moral and behavioral defects, and the Seven Deadly Sins.

Table 4.4: Significant Differences in HURTLEX Features for Sexist vs Non-Sexist Spanish TikToks. Bold values indicate the group with the highest mean for each category.

Category	**p-value**	**Non-sexist Mean**	**Non-sexist SD**	**Sexist Mean**	**Sexist SD**
an	0.0113	0.7481	2.9656	**0.8839**	2.6896
pr	$9.91 \cdot 10^{-5}$	0.3544	2.6003	**0.4372**	2.0165
re	$5.29 \cdot 10^{-11}$	0.3177	1.2594	**0.8401**	2.2798
dmc	0.0022	**0.7641**	3.1787	0.7598	2.2181
asf	$1.08 \cdot 10^{-7}$	0.3962	1.6483	**0.7889**	2.3363
qas	0.0234	1.0024	2.7009	**1.0426**	2.3437
om	0.0011	0.1387	0.8520	**0.3663**	2.5668
asm	0.0021	0.5880	1.9740	**0.8285**	2.2177
ddp	0.0004	0.5132	1.9032	**0.7731**	3.0536
or	0.0132	0.1586	0.8977	**0.3750**	2.5953
pa	0.0028	0.1424	1.0311	**0.1841**	1.1747

Table 4.5: Significant Differences in HURTLEX Features between Reported and Directed Sexism Spanish TikToks. Bold values indicate the group with the highest mean for each category.

Category	**p-value**	**Reported Mean**	**Reported SD**	**Direct Mean**	**Direct SD**
cds	0.0106	3.4295	4.4098	**3.3627**	5.5670
pr	0.0043	0.1661	0.9030	**0.5615**	2.3473
is	0.0248	**0.0906**	0.4616	0.0827	0.6434
re	3.77×10^{-14}	**1.2984**	2.5405	0.6301	2.1186
dmc	0.0166	0.7541	2.1284	**0.7625**	2.2594
ddp	0.0021	**0.8092**	1.9132	0.7566	3.4535
svp	0.0167	0.2478	1.1826	**0.2643**	1.5501

4.1.2.3 Emotions

To extract sentiments from the TikToks in Spanish, EmoRoBERTa was employed. EmoRoBERTa[6] is a model trained on the "GoEmotions" dataset[7], which consists of 58000 Reddit comments labeled with 28 different emotions, including admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief,

[6] https://huggingface.co/arpanghoshal/EmoRoBERTa

[7] https://huggingface.co/datasets/google-research-datasets/go_emotions

joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, and neutral. The use of EmoRoBERTa allows for a nuanced understanding of the emotions conveyed in the TikToks, capturing a wide range of emotional nuances.

The sentiment extraction process involved several steps. Firstly, the TikTok content, including the title, transcription, and OCR text, was translated from Spanish to English using the Googletrans library[8]. This translation step was crucial to ensure that the text could be processed by EmoRoBERTa, which is trained on English data. Subsequently, EmoRoBERTa was used to extract sentiments from the translated TikTok content, providing insights into the emotional undertones of the videos.

The emotional analysis of TikToks reveals distinct patterns based on their content related to sexism, as illustrated in Figure 4.3. Non-sexist videos tend to evoke positive emotions such as *amusement*, *desire*, *excitement*, *joy*, and *love*, while sexist TikToks show elevated levels of negative emotions including *anger*, *disapproval*, *disgust*, and *sadness*. Direct sexist TikToks primarily elicit *amusement*, *nervousness*, *relief*, and *neutral* emotions, possibly due to their shock value or perceived humor. In contrast, TikToks reporting sexism provoke stronger negative emotions like *disapproval*, *disgust*, *fear*, and *grief*, reflecting a critical and defensive stance towards the content, suggesting an active recognition and rejection of sexist undertones. For more details on emotional analysis, please refer to the following tables in the appendix A.6 and A.7.

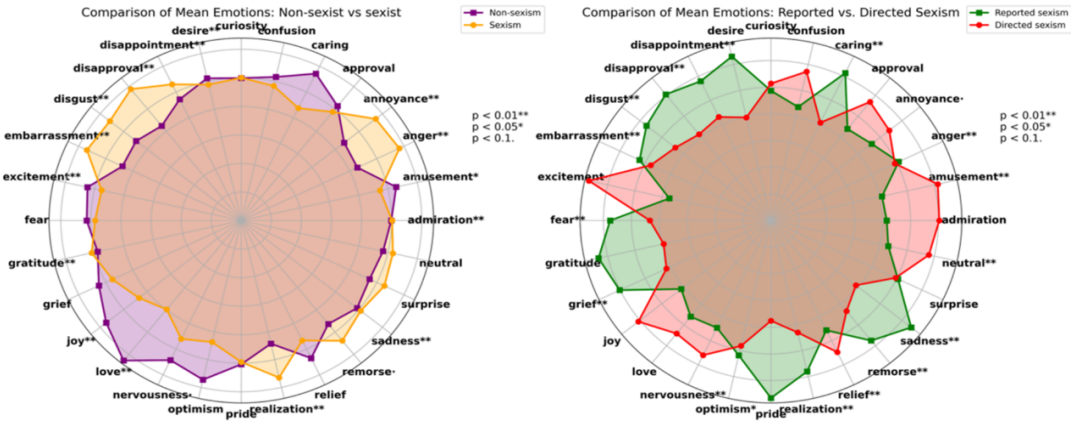


Figure 4.3: Radar chart comparing means of different emotions for EmoRoBERTa TikToks in Spanish (Non-sexist vs Sexist on the left and Reported vs Direct on the right).

4.1.2.4 BETO Contextualized Hate Speech

To extract scores related to hate speech variables, the *beto-contextualized-hate-speech* model[9] was employed. This model is trained to detect hate speech comments in news articles using a multilabel classification approach. The base model is BETO, a Spanish BERT pre-trained model, and the classification labels include categories such as WOMEN (against women), LGBTI (against LGBTI), RACISM (racist), CLASS (classist), POLITICS (because of politics), DISABLED (against disabled), APPEARANCE (against people because of their appearance), CRIMINAL (against criminals), and an extra label CALLS, representing whether a comment is a call to violent action or not.

The results presented in Table 4.6 show significant differences in hate speech scores between sexist and non-sexist Spanish TikToks. Overall, sexist TikToks have higher hate speech scores for all categories in the table except *CALLS*. Specifically, the *WOMEN* category exhibited the most notable differences.

When comparing reported and direct sexism, significant differences in hate speech scores were also evident as indicated in Table 4.7. In this case, directed sexism had higher scores for all categories in the table except *LGBTI*.

[8] https://pypi.org/project/googletrans/
[9] https://huggingface.co/piuba-bigdata/beto-contextualized-hate-speech

Table 4.6: Significant Differences in Hate Speech for Sexist vs Non-Sexist Spanish TikToks. Bold values indicate the group with the highest mean for each category.

Category	p-value	Non-sexist Mean	Non-sexist SD	Sexist Mean	Sexist SD
CALLS	3.89×10^{-17}	**0.0017**	0.0244	0.0011	0.0028
WOMEN	1.88×10^{-39}	0.0253	0.1126	**0.0766**	0.1948
LGBTI	6.79×10^{-21}	0.0094	0.0472	**0.0215**	0.0818
RACISM	0.0111	0.0046	0.0407	**0.0053**	0.0485
CLASS	1.14×10^{-16}	0.0021	0.0183	**0.0022**	0.0094
POLITICS	1.85×10^{-16}	0.0026	0.0319	**0.0034**	0.0332
DISABLED	4.80×10^{-16}	0.0024	0.0258	**0.0035**	0.0341
APPEARANCE	8.25×10^{-8}	0.0106	0.0788	**0.0157**	0.0982
CRIMINAL	1.30×10^{-27}	0.0013	0.0051	**0.0020**	0.0084

Table 4.7: Significant Differences in Hate Speech between Reported and Directed Sexism Spanish TikToks. Bold values indicate the group with the highest mean for each category.

Category	p-value	Reported Mean	Reported SD	Direct Mean	Direct SD
WOMEN	3.96×10^{-9}	0.0496	0.1591	**0.0890**	0.2081
LGBTI	2.12×10^{-7}	**0.0236**	0.0914	0.0205	0.0770
RACISM	0.0002	0.0020	0.0078	**0.0068**	0.0583
CLASS	2.45×10^{-11}	0.0013	0.0045	**0.0027**	0.0109
POLITICS	7.70×10^{-5}	0.0013	0.0021	**0.0044**	0.0401
DISABLED	1.36×10^{-9}	0.0015	0.0035	**0.0045**	0.0411
APPEARANCE	4.17×10^{-14}	0.0080	0.0630	**0.0192**	0.1106

4.1.2.5 Linguistic Resources for Sexism Categorization.

In this section, we explore the significant differences in linguistic features across various categories of sexism, as illustrated in Figure 4.4. This analysis highlights how sexism manifests through specific language use, offering insights crucial for developing effective categorization and detection systems on social media platforms like TikTok.

For the category of Ideological Inequality, the linguistic analysis reveals an increased usage of prepositions and terms associated with physical contact (MecCog), as well as a prevalence of inclusive terms such as 'union', 'encompass*', 'inside', 'included', 'including', and 'sum*'. Additionally, this category features a variety of social terms like 'homeland', 'donate', 'converse', 'manifesting', and 'fight'. These linguistic patterns suggest discussions that focus on social cohesion, collective identities, and challenging or reinforcing societal structures.

In contrast, the Stereotyping-Dominance category is characterized by a higher occurrence of household-related terms such as 'room*', 'curtain*', 'live', 'TV', and 'oven', with fewer sexual terms, indicating content that perpetuates traditional domestic roles and gender stereotypes within the household setting.

Content labeled as Objectification is notably marked by a higher frequency of profane words (Maldec) or higher scores of hate speech, particularly against people because of their appearance. This category generally avoids discussions on mortality, focusing instead on derogatory or demeaning language that degrades physical or aesthetic attributes.

The Sexual Violence category exhibits enhanced negative emotions and anger, with a predominant use of past tense verbs. This linguistic style is indicative of content that often recounts past incidents of violence, emphasizing the emotional and psychological trauma associated with such experiences.

Furthermore, posts identified as Misogyny are distinguished by high levels of negative emotions and frequent use of terms related to death. These patterns reflect extreme hostility or disdain towards women, possibly conveying a dehumanizing perspective or an existential threat.

By dissecting these linguistic distinctions, we gain valuable insights into how sexism is contextually embedded in language. This understanding is instrumental for the development of algorithms capable of detecting nuanced and context-specific instances of sexism, enhancing both the accuracy and effectiveness of monitoring and intervention strategies on digital platforms. This linguistic approach not only aids in the precise categorization of sexist content but also enriches our comprehension of the complex psychological and social dynamics that fuel online expressions of sexism.

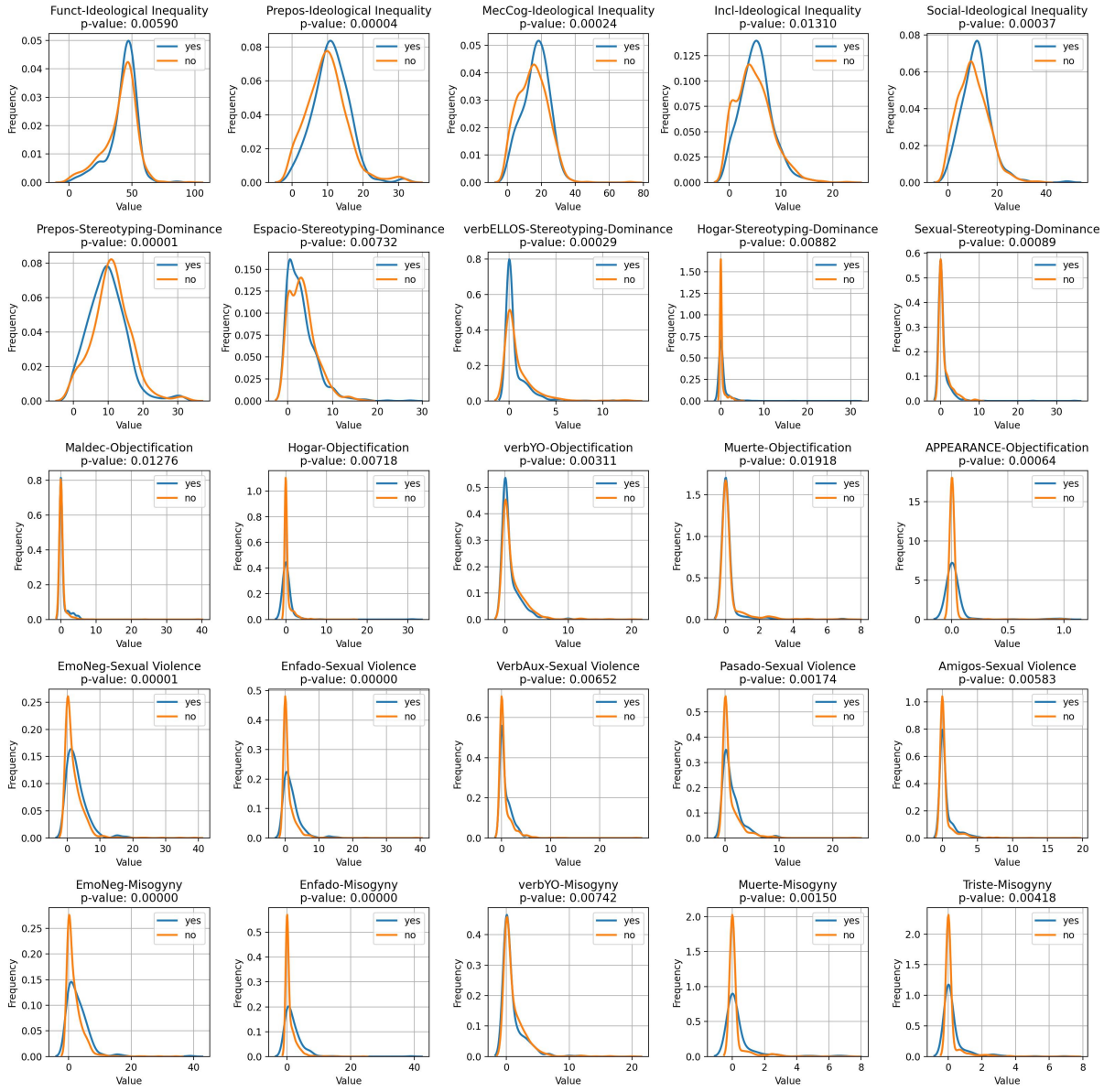


Figure 4.4: Examples of Distribution of Significant Linguistic Resource Features for Sexism Categories,

4.1.3 Feature Representation

In this section, we will discuss how the transcriptions, OCR, and titles of the TikToks will be transformed into different representations, which will be the features that the models will receive for the three tasks: determining whether a TikTok is sexist or not, determining the intention of the TikTok (direct or reported), and classifying the different categories of sexism.

4.1.3.1 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a widely used statistical method in natural language processing and information retrieval. It is used to measure the importance of a term within a document relative to a collection of documents (i.e., relative to a corpus). The TF-IDF representation is calculated as follows:

- **Term Frequency (TF)**: It is the frequency of a term or word in a document, divided by the total number of words in the document.

$$\text{TF}(t, d) = \frac{\text{number of times term } t \text{ appears in document } d}{\text{total number of terms in document } d}$$

- **Inverse Document Frequency (IDF)**: It reflects the proportion of documents in the corpus that contain the term. Terms unique to a small percentage of documents receive higher importance values than words common across all documents.

$$\text{IDF}(t, D) = \log\left(\frac{\text{total number of documents in } D}{\text{number of documents with term } t + 1}\right)$$

- **TF-IDF**: It is the product of TF and IDF.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

This representation will be applied to both English and Spanish TikToks.

4.1.3.2 Pretrained Transformers

Pretrained transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (A Robustly Optimized BERT Pretraining Approach), have proven to be highly effective in a variety of natural language processing tasks. These models learn an internal representation of language that can be used to extract useful features for downstream tasks. One of the most popular pretraining techniques used in these models are:

- **Masked Language Modeling (MLM)**: In this approach, some words in a sentence are masked, and the model has to predict them based on the provided context.
- **Next Sentence Prediction (NSP)**: The model learns to predict whether a sentence is the next one in a given pair of sentences.

These representations can be used directly as features for machine learning models. Moreover, pretrained models can be fine-tuned for specific tasks, adapting the features learned during pretraining to the particular task at hand. The transformers we are going to use in both languages are:

English:

- **FacebookAI/roberta-large**[10]: RoBERTa is a transformer-based model pretrained on a large corpus of English data in a self-supervised manner. This means that it was pretrained on raw texts only, without any human labeling, allowing it to utilize a large amount of publicly available data.
- **cardiffnlp/twitter-roberta-base-hate-multiclass-latest**[11]: This model is a fine-tuned version of cardiffnlp/twitter-roberta-base-2022-154m for multiclass hate-speech classification. A combination of 13 different hate-speech datasets in the English language was used to fine-tune the model.

[10] `https://huggingface.co/FacebookAI/roberta-large`
[11] `https://huggingface.co/cardiffnlp/twitter-roberta-base-hate-multiclass-latest`

Spanish:

- **PlanTL-GOB-ES/roberta-large-bne**[12]: The roberta-large-bne is a transformer-based masked language model for the Spanish language. It is based on the RoBERTa large model and has been pre-trained using the largest Spanish corpus known to date, with a total of 570GB of clean and deduplicated text processed for this work, compiled from the web crawlings performed by the National Library of Spain (Biblioteca Nacional de España) from 2009 to 2019[13].
- **piuba-bigdata/beto-contextualized-hate-speech**: This model has been previously explained in the Linguistic Resources section for extracting hate speech variables.

In summary, these representations and pretrained models provide a variety of features and approaches to address the proposed tasks, offering greater flexibility and potential for obtaining accurate and robust results.

4.1.4 Machine Leaning Models

In this section, we will introduce three machine learning models for the three classification tasks related to TikToks: sexism detection, source intention classification, and sexism categorization. These models will receive two types of features. Firstly, textual features extracted from the title, transcription, and OCR of the TikToks, represented using TF-IDF or embeddings. Secondly, linguistic features derived from resources such as LIWC, HURTLEX, emotion scores, and hate speech detection will be incorporated. By integrating these features, the models aim to provide a comprehensive and nuanced analysis of TikTok content, facilitating accurate identification and classification of various forms of content and intentions.

4.1.4.1 Support Vector Machine

Support Vector Machine (SVM)[14] is a supervised machine learning algorithm used for classification and regression tasks. It is particularly effective in high-dimensional spaces and is well-suited for both linear and nonlinear classification. Figure 4.5 provides a visualization of a Linear SVM on 2D data, highlighting the separation achieved by the SVM hyperplane.

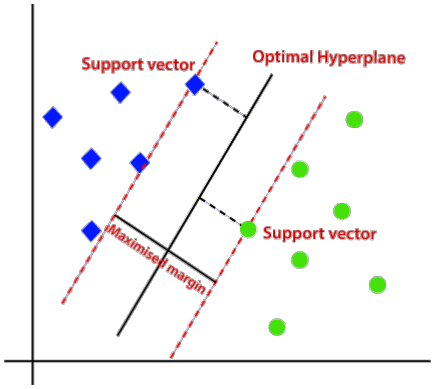


Figure 4.5: Visualization of Linear SVM on 2D Data

Linear SVM Classification. Linear SVM classification aims to find the best hyperplane that separates the data points of different classes with the largest margin. The margin is defined as the distance between the hyperplane and the nearest data point of any class. The objective is to maximize this margin, which helps in achieving better generalization to unseen data. There are two types of margins:

- **Hard Margin**. In hard margin SVM, the algorithm strictly enforces that all data points are correctly classified and are located outside the margin boundaries. This approach can be sensitive to outliers and may not be suitable for datasets that are not linearly separable.

[12] https://www.bne.es/en
[13] https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne
[14] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

- **Soft Margin**. Soft margin SVM allows some data points to be misclassified and to be located within the margin boundaries. The parameter C controls the trade-off between maximizing the margin and minimizing the classification error. A smaller value of C allows for a wider margin but may result in more margin violations, while a larger value of C results in a narrower margin with fewer margin violations. Figure 4.6 illustrates the difference between large margin (left) and fewer margin violations (right).

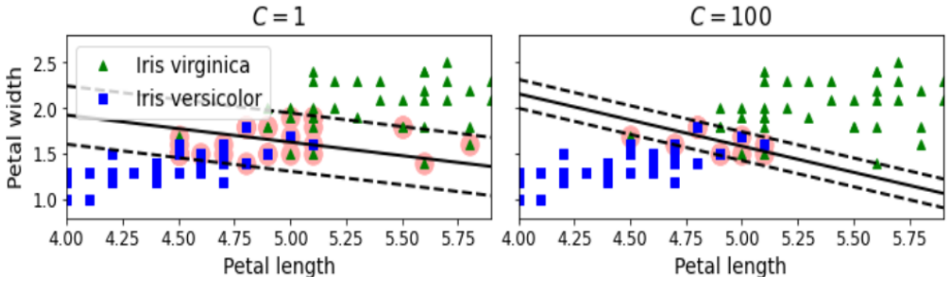


Figure 4.6: Large margin (left) vs fewer margin violations (right)

Nonlinear SVM Classification. Linear SVM is effective when the data is linearly separable. However, many real-world datasets are not linearly separable. Nonlinear SVM classification addresses this issue by mapping the original features into a higher-dimensional space where the data becomes linearly separable.

The **Kernel Trick** is a technique used in SVM to efficiently compute the dot products in the higher-dimensional space without explicitly transforming the data. One popular kernel is the **Gaussian Radial Basis Function (RBF)** kernel, that is defined as:

$$K(\mathbf{a}, \mathbf{b}) = \exp\left(-\gamma \|\mathbf{a} - \mathbf{b}\|^2\right)$$

where γ is a hyperparameter that controls the spread of the kernel. The Gaussian RBF kernel allows SVM to capture complex nonlinear relationships in the data by measuring the similarity between data points in the original feature space.

4.1.4.2 Multilayer Perceptron

The Multilayer Perceptron (MLP)[15] is an artificial neural network architecture consisting of multiple layers of nodes or neurons, including an input layer, one or more hidden layers, and an output layer. It is one of the most widely used and versatile neural networks, capable of learning and modeling complex functions.

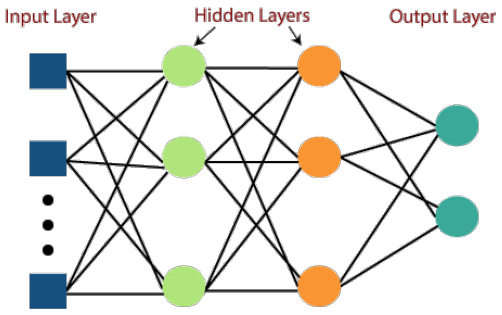


Figure 4.7: Structure of a Multilayer Perceptron (MLP)

The MLP uses the backpropagation algorithm to train the neural network and adjust the weights of the connections between the neurons. The objective during training is to minimize a loss function, such as the Mean Squared Error (MSE) for regression problems or the Cross-Entropy for classification problems. Some of the key hyperparameters in training an MLP include:

[15] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

- **Number of Hidden Layers and Neurons**. Determines the capacity and complexity of the model. A higher number of layers and neurons can lead to overfitting if not properly regularized.
- **Activation Functions**. Choosing the appropriate activation function for the hidden and output layers, such as the sigmoid function, ReLU, hyperbolic tangent, etc.
- **Learning Rate**. Controls the magnitude of the weight adjustments during training. A too high learning rate may cause the model not to converge, while a too low learning rate may make the training too slow.
- **Regularization and Dropout**. Techniques used to prevent overfitting, such as L1/L2 regularization and dropout.

In conclusion, the Multilayer Perceptron is a powerful artificial neural network architecture that offers flexibility and the ability to model complex functions. With proper hyperparameter selection and regularization techniques, the MLP can be trained to achieve good performance on a wide variety of machine learning tasks.

4.1.4.3 Extra-Trees: An Extremely Randomized Trees Ensemble Technique

In Extremely Randomized Trees (Extra-Trees)[16], the level of randomness is significantly increased during the split selection process compared to traditional Random Forests (RF). Unlike Random Forests, which utilize bootstrapping to create multiple subsets of the original dataset for training individual trees, Extra-Trees train each tree on the entire original dataset.

While both Random Forests and Extra-Trees consider a random subset of features for making splits, Extra-Trees take the randomness a step further. For each candidate feature in Extra-Trees, multiple thresholds are chosen randomly, and the best among these randomly selected thresholds is picked as the splitting rule. This additional level of randomness can lead to a slight increase in bias but often results in a significant reduction in the model's variance.

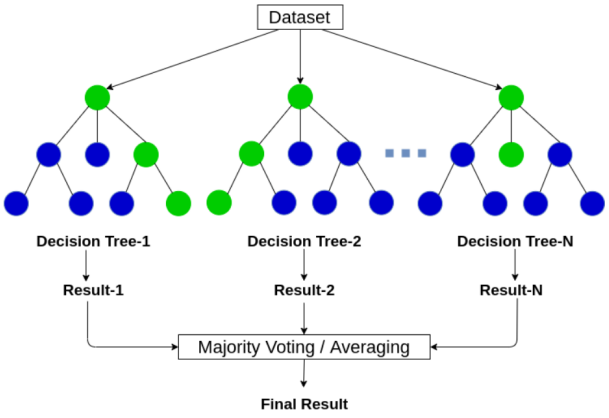


Figure 4.8: Ensemble of decision trees

Classification

For classification tasks in Extra-Trees, the objective is to minimize either the Gini impurity or the entropy to maximize the information gain. The information gain, represented as $\Delta H(S_n, s)$, is calculated using the following formula:

$$\Delta H(S_n, s) = H(S_n) - \left(\frac{|S_l|}{|S_n|} \cdot H(S_l) + \frac{|S_r|}{|S_n|} \cdot H(S_r) \right) \tag{4.1}$$

Where:

- $H(S_n)$ is the entropy at node n, calculated as $H(S_n) = -\sum_{i=1}^{c} p(i|S_n) \log_2 p(i|S_n)$, where $p(i|S_n)$ is the proportion of class i samples in node n.

[16] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

- $|S_l|$ and $|S_r|$ are the sizes of the left and right child nodes S_l and S_r respectively.
- $H(S_l)$ and $H(S_r)$ are the entropies of the left and right child nodes S_l and S_r respectively.

Here, s represents the split defined by the pair (kn, vn), where kn is the randomly selected feature, and vn is the randomly selected threshold for that feature.

The increased level of randomization in Extra-Trees offers several advantages, including reduced variance, decreased overfitting, and improved computational efficiency. By training each tree on the full dataset and using random splits, Extra-Trees can effectively capture intricate patterns in the data, resulting in robust performance across a wide range of machine learning tasks.

Important Hyperparameters

- **Number of Trees** (*n_estimators*): Specifies the number of trees in the ensemble. A higher number of trees can improve performance but also increases computational cost.
- **Maximum Features** (*max_features*): Determines the number of features to consider when looking for the best split. It can be a fixed number, a percentage, or 'auto'/'sqrt' (square root of the total number of features).
- **Minimum Samples Split** (*min_samples_split*): The minimum number of samples required to split an internal node. Increasing this value can lead to a more robust model by preventing overfitting.
- **Maximum Depth** (*max_depth*): The maximum depth of the tree. It limits the depth of individual trees, helping to prevent overfitting.

4.1.4.4 Stacking Classifier

The Stacking Classifier[17] is an ensemble learning technique that combines multiple classification models to improve overall performance and predictive accuracy. In the Stacking Classifier setup, a stack of diverse base estimators is used in conjunction with a meta-classifier, typically Logistic Regression, to produce the final prediction. Figure 5.3 illustrates the concept of the Stacking Classifier and its components.

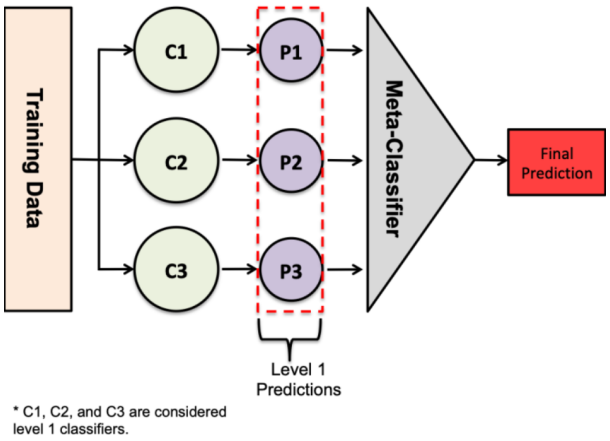


Figure 4.9: Illustration of the Stacking Classifier

Here's a breakdown of how the Stacking Classifier operates:

1. **Base Estimators.** Several different classification algorithms, such as SVM, MLP, and Extra-Trees, are trained on the dataset. Each base estimator generates its own set of predictions based on the input data.
2. **Stacking.** The predictions from these base estimators are utilized as new features. Rather than using the original dataset features, the Stacking Classifier employs the predictions of the base estimators as its inputs.

[17] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html`

3. **Meta-Classifier.** A final classifier, typically Logistic Regression, is trained on the new set of features, which are the predictions from the base estimators. This meta-classifier learns to combine the predictions of the base estimators to produce the ultimate prediction.

The primary concept behind Stacking is to harness the strengths of individual base estimators. Each base estimator might excel at capturing specific patterns or characteristics of the data. By aggregating their predictions through a meta-classifier, the Stacking Classifier aims to achieve superior generalization and enhanced predictive accuracy compared to using any single base estimator in isolation.

4.2 Audio Models

4.2.1 Feature Representation

Feature representation is a crucial step in audio signal processing, where raw audio signals are transformed into a format suitable for machine learning algorithms. This process involves extracting relevant features that capture essential characteristics of the audio signals, enabling effective analysis and classification tasks.

4.2.1.1 Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCC) are widely used in speech and audio processing due to their effectiveness in capturing the spectral characteristics of audio signals. The MFCC extraction process involves several steps:

1. **Pre-emphasis.** The raw audio signal is pre-emphasized to amplify high-frequency components, enhancing the signal-to-noise ratio.
2. **Frame Blocking**. The pre-emphasized signal is divided into short frames, typically around 20-40 milliseconds long, to capture temporal variations in the signal.
3. **Windowing**. Each frame is multiplied by a window function, such as the Hamming or Hanning window, to reduce spectral leakage and smoothen the edges of the frame.
4. **Fast Fourier Transform (FFT).** The windowed frames are passed through the FFT to convert the signal from the time domain to the frequency domain, resulting in a spectrum representation.
5. **Mel Filterbank.** The spectrum is passed through a series of Mel filters, which mimic the non-linear frequency response of the human auditory system. These filters are spaced uniformly in the Mel scale to better capture human perception of sound.
6. **Log Compression.** The logarithm of the filterbank energies is computed to compress the dynamic range of the spectrum and make the features more robust to variations in signal intensity.
7. **Discrete Cosine Transform (DCT).** Finally, the DCT is applied to decorrelate the MFCC coefficients, resulting in a compact representation of the spectral features.

The resulting MFCCs represent the spectral envelope of the audio signal, capturing important characteristics such as timbre and pitch. These coefficients are commonly used as features in various audio processing tasks, including speech recognition, speaker identification, and music genre classification.

In the context of detecting sexism in TikTok videos, MFCCs can play a crucial role. Since audio is a significant component of TikTok content, analyzing the audio features, such as speech patterns and intonations, can provide valuable insights into the underlying sentiments and tones of the videos. MFCCs, being effective representations of audio signals, can help in extracting discriminatory language patterns or identifying specific linguistic cues associated with sexist content.

In addition to Mel-Frequency Cepstral Coefficients (MFCCs), several other audio features will be extracted to provide a comprehensive representation of the audio signal:

- **MFCCs**. Extraction of 40 MFCC coefficients is planned. In practice, the first 8–13 MFCC coefficients are commonly used to represent the shape of the spectrum. However, some applications require higher-order coefficients to capture additional information such as pitch and tone nuances.

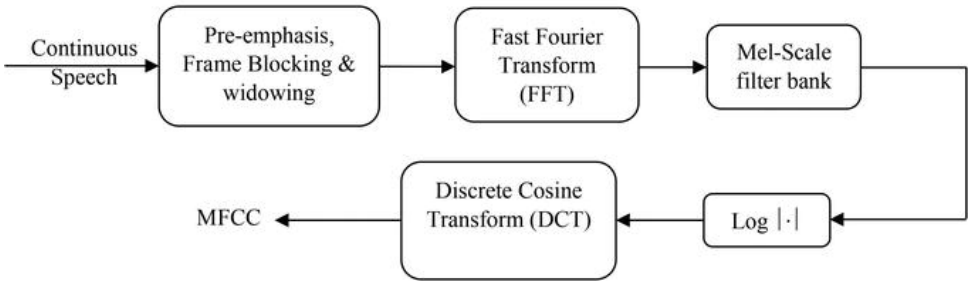


Figure 4.10: Overview of the MFCC extraction process.

- **Chroma**. Chroma features represent the energy distribution of pitch classes (or musical notes) in an audio signal, providing information about its tonal content. There are 12 chroma features, denoted as `chroma_1` to `chroma_12`.
- **Spectral Contrast**. Measurement of the difference in amplitude between peaks and valleys in the spectrum, providing information about the spectral texture or brightness of the audio signal. There are 7 spectral contrast features, denoted as `spectral_contrast_1` to `spectral_contrast_7`.
- **Tonnetz**. Tonnetz features represent tonal centroids in a pitch class space, capturing tonal characteristics such as harmonic and melodic content. There are 6 tonnetz features, denoted as `tonnetz_1` to `tonnetz_6`.
- **Root Mean Square (RMS)**. RMS is a measure of the overall energy level of an audio signal, providing information about its amplitude distribution. There is 1 RMS feature.
- **Zero Crossing Rate (ZCR)**. ZCR measures the rate at which the audio signal changes sign (crosses the zero amplitude line), offering insights into the signal's temporal dynamics. There is 1 ZCR feature.

In total, 67 features will be extracted, including the 40 MFCC coefficients and additional features such as chroma, spectral contrast, tonnetz, RMS, and ZCR. These features collectively capture various aspects of the audio signal, enabling a comprehensive analysis for tasks such as classification, clustering, and discrimination detection. It is important to note that although some examples of these feature values are shown across individual frames, classifiers typically work with aggregate statistics, such as the mean, across all frames.

4.2.1.2 Pre-trained Wav2Vec2 Embeddings

The Wav2Vec2 model, proposed in [5], presents a novel approach to feature extraction from audio signals. This model demonstrates the effectiveness of learning powerful representations from speech audio alone, followed by fine-tuning on transcribed speech data.

For feature extraction, pre-trained models such as `jonatasgrosman/wav2vec2-large-xlsr-53-spanish`[18] for Spanish and `jonatasgrosman/wav2vec2-large-xlsr-53-english`[19] for English have been utilized. These models are pre-trained on the Common Voice 6.1 dataset, which includes a vast amount of recorded hours of speech data along with corresponding text files. The dataset also contains demographic metadata such as age, sex, and accent, which can aid in training the accuracy of speech recognition engines.

The model architecture comprises a multi-layer convolutional feature encoder $f : X \rightarrow Z$, which takes raw audio X as input and generates latent speech representations $z_1, \ldots, z_T$ for T time-steps. These representations are then processed by a Transformer $g : Z \rightarrow C$ to produce representations $c_1, \ldots, c_T$ capturing information from the entire sequence. The output of the feature encoder is discretized to q_t using a quantization module $Z \rightarrow Q_t$ to represent the targets in the self-supervised objective. This architecture is represented in Figure 4.11.

[18] `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-spanish`
[19] `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`

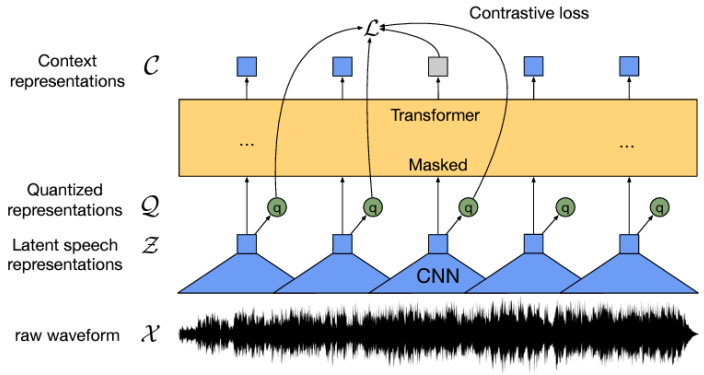


Figure 4.11: Wav2Vec2 Architecture.

4.2.2 Machine Leaning Models

In the context of audio models for analyzing TikTok content, the same machine learning models described earlier for text classification tasks are utilized. MFCCs or embeddings from Wav2Vec2 serve as inputs to these classifiers. These features capture essential characteristics of the audio signals, providing a robust foundation for the models to perform tasks like detecting sexist content, classifying sources of intention, and categorizing types of sexism in TikTok videos

4.3 Video Models

While text and audio analysis are fundamental in detecting sexist content, video models offer a unique perspective by capturing visual features and temporal dynamics that may go unnoticed in other media. Gestures, facial expressions, visual environments, and video edits can provide important clues about the tone and intent of a video, thus complementing text and audio-based analysis.

In this study, we investigate three distinct video models tailored for the purpose of identifying instances of sexism within TikTok content. These models employ varied methodologies, ranging from the extraction of visual features through convolutional neural networks or transformers to the generation of frame-level descriptions. By combining these models with text and audio analysis, we can gain a more comprehensive understanding of TikTok content and develop more effective tools for identifying and addressing sexism on this platform.

4.3.1 ResNet+LSTM

This model extracts N frames from TikTok videos. Each frame undergoes feature extraction using a pre-trained ResNet model, which has been pre-trained on ImageNet, comprising 14 million images and 24 million parameters. The ResNet extracts a 2048-dimensional feature vector for each frame. These feature vectors serve as inputs to a Long Short-Term Memory (LSTM) network, which models the temporal relationships between the frames. The output of the LSTM passes through several fully connected layers with ReLU activation and Dropout regularization. Finally, the model produces a probability indicating whether the TikTok video is sexist (for Task 1), the intention is direct sexism (for Task 2), or a probability distribution across different categories of sexism (for Task 3). Figure 4.12 illustrates the architecture of the ResNet+LSTM model.

4.3.2 ViT+LSTM

Similar to the previous model, this one extracts features from each frame, but instead of using Convolutional Neural Networks (CNNs) like ResNet, it utilizes Vision Transformers (ViT). ViT extracts 768-dimensional feature vectors for each frame, which are then temporally modeled by an LSTM. The LSTM output is fed through fully connected layers to produce the final prediction. Figure 4.13 depicts the architecture of the ViT+LSTM model.

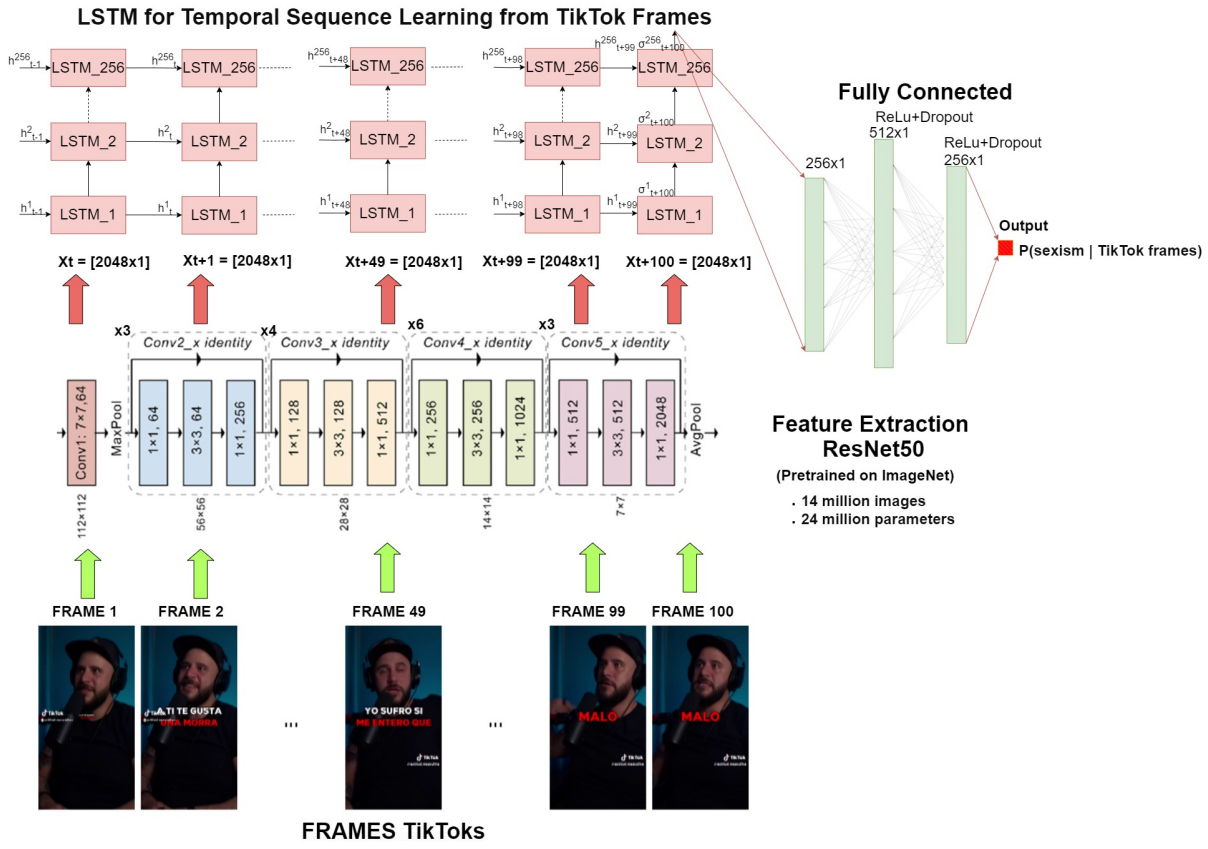


Figure 4.12: Architecture of ResNet+LSTM model.

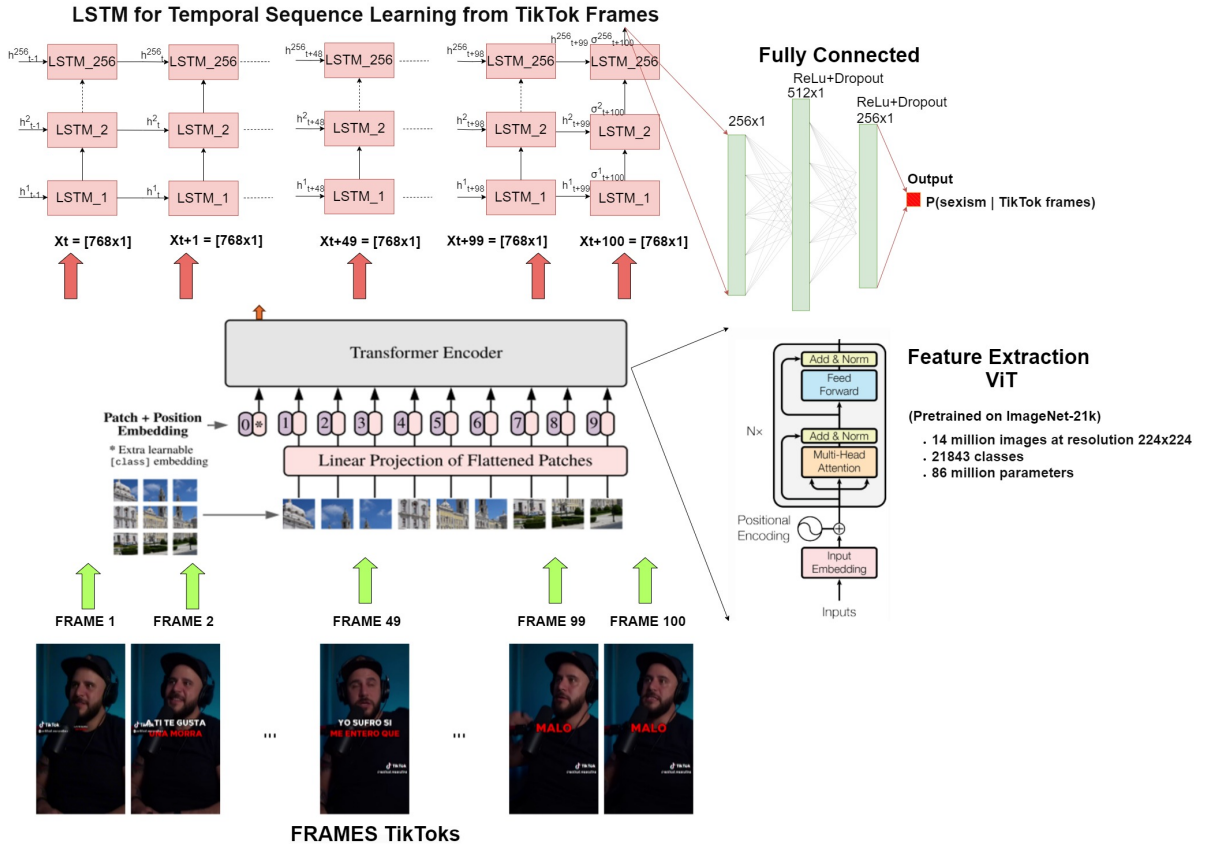


Figure 4.13: Architecture of ViT+LSTM model.

4.3.3 BLIP+TF-IDF

In this model, N frames from each TikTok video are processed to generate captions using the `unography/blip-large-long-cap`[20] model, which has been fine-tuned on the `unography/laion-14k-GPT4V-LIVIS-Captions` dataset[21]. Each caption undergoes TF-IDF transformation to create feature vectors, which are then processed by fully connected layers or any other machine learning model for the final prediction. Figure 4.14 showcases the architecture of the BLIP+TF-IDF model.

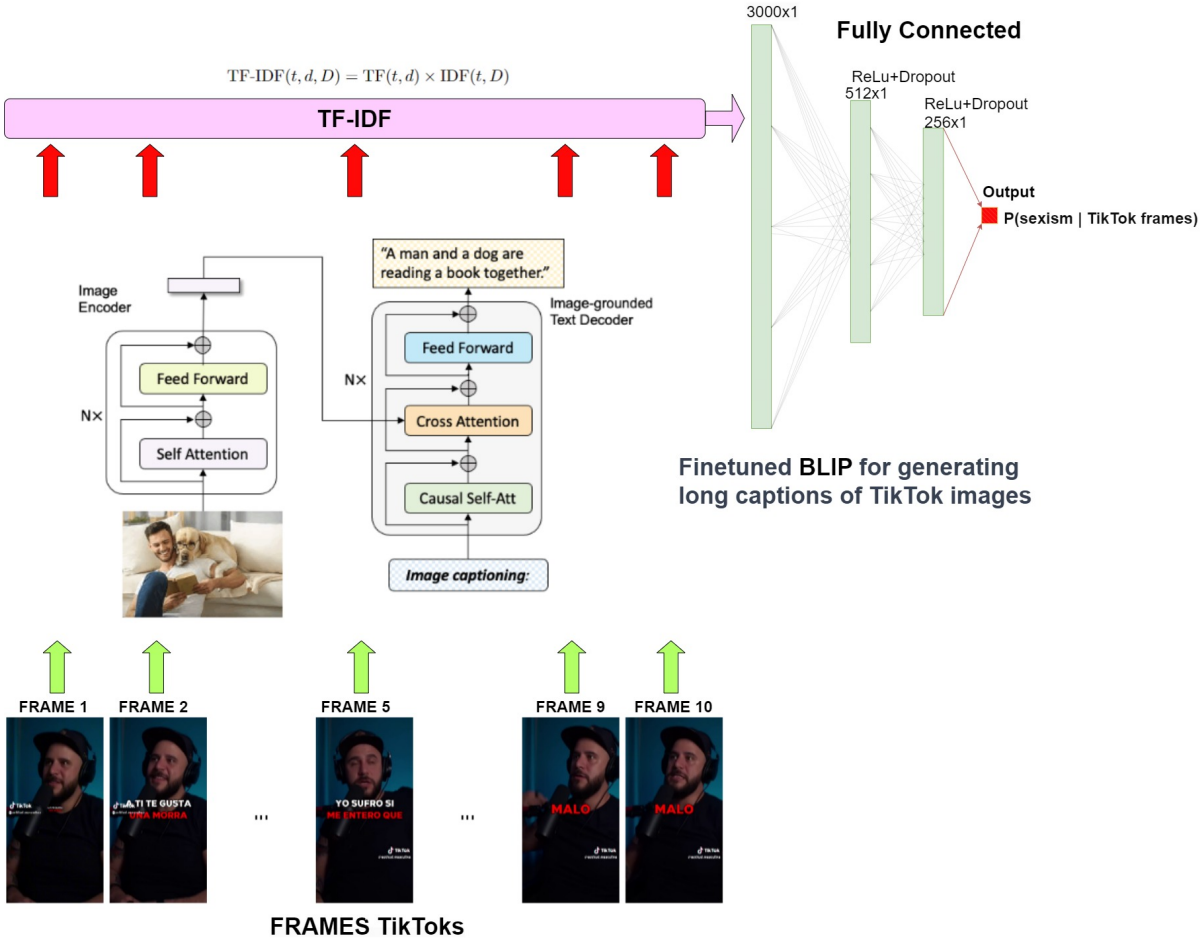


Figure 4.14: Architecture of BLIP+TF-IDF model.

These models collectively provide different approaches to analyzing video content for detecting sexism in TikTok videos. Each model leverages distinct feature extraction methods and architectures to capture relevant information from the video data.

4.4 Multimodal Models

4.4.1 Multimodal SVM

Our comprehensive approach to detect sexism in TikTok videos involves an SVM that integrates all modalities: text, audio, and video. The text component is represented by embeddings; the audio by MFCCs or Wav2Vec2 embeddings; and the video by average pooled features from ResNet, ViT, or TF-IDF vectors derived from captions generated by the BLIP model. This choice was driven by the limited size of our dataset, which made the average pooling of video features more effective than sequential processing with LSTMs. Ablation tests were conducted to determine the impact of each modality on the overall performance of the model. By excluding one type of feature (text, audio, or video) at a time,

[20] `https://huggingface.co/unography/blip-large-long-cap`
[21] `https://huggingface.co/datasets/unography/laion-14k-GPT4V-LIVIS-Captions`

we were able to identify the critical components necessary for effective sexism detection. These tests highlight the unique contributions of each modality.

4.4.2 Text, Audio, Video and Linguistic Features (TAVL) - Fine-Tuning

Additionally, we experimented with creating a multimodal model that combines textual, audio, and video information from the TikToks, not only by concatenating pre-trained embeddings but also by allowing the fine-tuning of the last attention layers of each model to better adjust to our specific task of sexism detection. The advantage of this approach is that it permits the adjustment of the attention layers considering all modalities, enhancing the model's ability to understand and integrate the various types of input data effectively. For the visual component, it employs the Vision Transformer (ViT), which processes video frames, followed by an LSTM layer to capture temporal features, followed by a dense layer. For the audio component, it uses Wav2Vec2 to extract embeddings from audio inputs, which are then passed through a dense layer. For textual data, the model utilizes a RoBERTa-based model to generate text embeddings, followed by a dense layer. Additionally, linguistic features (LIWC, HurtLex, hate speech scores, and emotion scores) are processed through a linear layer. The outputs from these components are concatenated and passed through a classifier, which consists of fully connected layers with dropout and ReLU activations to predict sexism. This architecture allows the model to fine-tune the attention layers specific to each modality while leveraging the pre-trained knowledge, ensuring a comprehensive and adaptive approach to sexism detection in multi-modal data. Figure 4.15 shows this architecture integrating multiple data modalities.

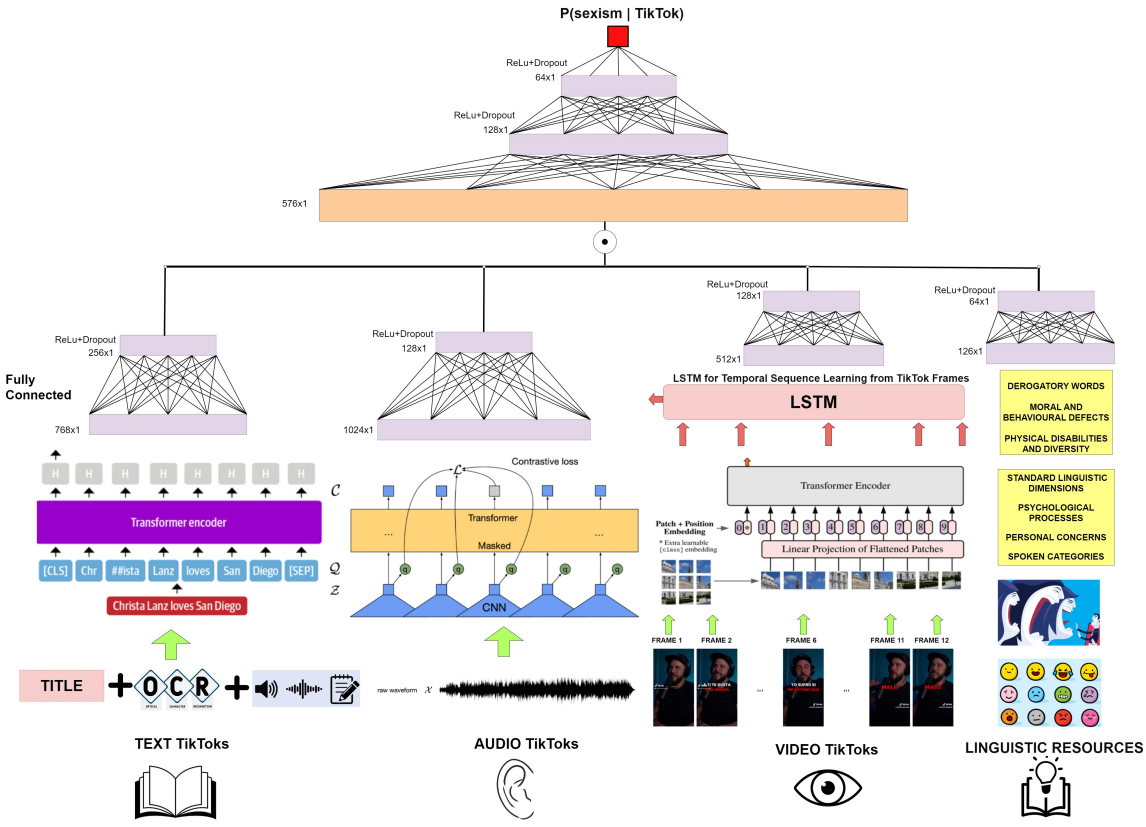


Figure 4.15: Architecture of the multimodal model

Table 4.8: Trainable Parameters for Each Block

Block	**Trainable Parameters**
ViT + Dense (vit.encoder.layer.6 to layer.11.*)	82,573,584
LSTM	1,051,648
Wav2Vec2 + Dense (wav2vec.encoder.layers.18 to layers.23.*)	74,846,976
RoBERTa + Dense (roberta.encoder.layer.6 to layer.11.*)	41,800,832
Linguistic Features	8,128
Final Classification	116,145
Total Trainable Parameters	**201,397,313**

Chapter 5

Text, Audio, Video, and Multimodal Experiments

5.1 Experimental Setup and Metric Explanation

5.1.1 Experimental Setup

The experimental setup entails employing robust techniques for model evaluation and hyperparameter optimization. To achieve this, a combination of 10-fold cross-validation and grid search methodology is utilized.

Cross-validation

Cross-validation is a widely adopted technique in machine learning for evaluating model performance. In 10-fold cross-validation, the dataset is divided into 10 equally sized folds. Iteratively, 9 folds are used for training the model, while the remaining fold is held out for validation. This process is repeated 10 times, with each fold serving as the validation set exactly once. By averaging the performance across these 10 iterations, a more reliable estimation of the model's generalization ability is obtained.

Grid Search

Grid search is a systematic approach to hyperparameter tuning, where a predefined set of hyperparameters is exhaustively searched to identify the combination that yields the best model performance. For each combination of hyperparameters, the model is trained and evaluated using cross-validation. The hyperparameters considered for tuning typically include regularization parameters, learning rates, and kernel sizes, among others. By exploring a grid of hyperparameter values, grid search helps identify the optimal configuration that maximizes the model's performance on the validation data.

By combining 10-fold cross-validation with grid search, the experimental setup ensures robust model evaluation and hyperparameter optimization, thereby enhancing the reliability and generalization capability of the trained models.

5.1.2 Evaluation Metric

The primary evaluation metric used for model selection across all three tasks is the F1-macro score. F1-macro considers both precision and recall, providing a balanced assessment of the model's performance across all classes. For each task, additional metrics such as F1 score, precision, and recall are computed, with a focus on the positive class relevant to the specific task.

- In the first task, the objective is to detect sexist content in TikTok videos. The positive class for this task, denoted as S, represents videos that contain sexist content. The model's performance is evaluated based on its ability to correctly classify videos as sexist or non-sexist. Precision measures the accuracy of identifying sexist videos among those predicted as such, while recall indicates the model's ability to capture all actual sexist videos among those present in the dataset.
- For the second task, the goal is to identify TikTok videos that contain directed sexism. The positive class, denoted as D, comprises videos that explicitly promote gender stereotypes or perpetuate sexist beliefs. Precision measures the accuracy of identifying direct sexist videos and recall indicates the model's ability to capture all instances of directed sexism among the videos.

- The third task involves categorizing various aspects of sexism present in TikTok videos. this task is multi-label, meaning each video may belong to multiple categories simultaneously. The model aims to classify videos into distinct categories such as ideological and inequality, role stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence. The evaluation metric used for this task is F1-macro, which calculates the average F1 score across all categories.

5.1.2.1 F1 Score

The F1 score is the harmonic mean of precision and recall, calculated as:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

5.1.2.2 Precision

Precision is the ratio of true positive predictions to the total number of positive predictions, calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

5.1.2.3 Recall

Recall, also known as sensitivity, is the ratio of true positive predictions to the total number of actual positives, calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

5.1.2.4 Macro F1 Score

Macro F1 score is the average of F1 scores for each class, giving equal weight to each class, calculated as:

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^{N} F1_i$$

where N is the number of classes.

5.2 Text Results

In the experiments utilizing textual features such as title, transcription, OCR data, several key comparisons were made to understand the impact of different feature representations and linguistic variables on model performance.

Firstly, the effectiveness of TF-IDF representation was compared to two types of pre-trained embeddings for English and two for Spanish. The TF-IDF representation was specifically configured with a fixed feature dimensionality of 3000 and considered up to trigrams to capture the textual information adequately.

Furthermore, the experiments explored whether augmenting textual features with linguistic variables, including LIWC, HURTLEX, emotion scores, and hate speech scores, improved the predictive capability of the models.

The results obtained from the experiments are presented in Tables 5.1 for English and 5.2 for Spanish. In general, we observe that the embeddings have surpassed the TF-IDF representation in terms of model performance. This suggests that the embeddings are better at capturing the underlying semantic relationships within the text data compared to the TF-IDF representation.

Interestingly, when we augment the textual features with linguistic variables, including LIWC, HURTLEX, emotion scores, and hate speech scores, we do not observe significant improvements in model performance when using embeddings. This phenomenon can be attributed to the fact that embeddings are already capable of capturing linguistic nuances and context, rendering the additional linguistic variables redundant.

However, when we add these linguistic variables to the TF-IDF representation, we see a notable improvement in model performance. This is likely because TF-IDF alone may not capture all the

subtle linguistic features present in the text data, and the linguistic variables provide complementary information that enhances the model's predictive capability.

For instance, Figure 5.1 illustrates the results of grid search cross-validation with TF-IDF on English TikToks for SVM and ExtraTrees. Here, we observe the influence of the parameter C and the number of trees on the macro F1 score for Task 2 of detecting sexism intention. We can see that adding linguistic variables consistently leads to an improvement compared to considering only the words and n-grams of TF-IDF.

More concretely, the best results in English for Task 1 were obtained with SVM (without using the extra linguistic variables) with pre-trained embeddings 2 (cardiffnlp/twitter-roberta-base-hate-multiclass-lates), achieving a macro F1 of 0.728 and an F1 for the sexist class of 0.705. For Task 2, the best model in terms of macro F1 was again an SVM without using the linguistic variables but with pre-trained embeddings 1 (FacebookAI/roberta-large) achieving 0.701. However, in terms of F1 for the direct class, the best model was an MLP without linguistic variables trained on these same embeddings, achieving 0.808. For the third task, the results in general were not good due to the lack of data for many of the sexism categories. The best result was obtained with an SVM with linguistic features on embeddings 2 with a macro F1 of 0.49.

For Spanish, the best results for Task 1 in terms of macro F1 were obtained with a stacking of models and with linguistic features trained on embeddings 2 (piuba-bigdata/beto-contextualized-hate-speech) with a macro F1 of 0.696. It is also the model that achieves the highest F1 for the sexist class with an F1 of 0.757. Surprisingly, for the second task, in this case, the Extra-Trees model with TF-IDF features wins in macro F1 with 0.696, while in F1 for the positive class, an MLP on embeddings 1 (PlanTL-GOB-ES/roberta-large-bne) wins with an F1 of 0.830. For the third task, the results are still poor, and the best model is an SVM with linguistic features on embeddings 1, with a macro F1 of 0.498.

Table 5.1: Results of the models across three tasks related to sexism on TikTok in the English corpus. 'S' refers to Sexist and 'D' denotes Direct sexism, which are the positive classes. Values in bold highlight the top-performing model for each metric across the tasks, while underlined values indicate the second highest performance.

Model	Task 1				Task 2				Task 3
	M-F1	F1 (S)	P (S)	R (S)	M-F1	F1 (D)	P (D)	R (D)	M-F1
TF-IDF									
SVM	0.671	0.611	0.680	0.558	0.666	0.787	0.731	0.854	0.323
SVM + ling.	0.679	0.635	0.661	0.614	0.678	0.779	0.751	0.812	0.343
MLP	0.641	0.602	0.614	0.598	0.661	0.786	0.729	0.858	0.294
MLP + ling.	0.658	0.621	0.630	0.617	0.669	0.782	0.740	0.832	0.300
Extra-Trees	0.698	0.658	0.685	0.634	0.657	0.759	0.736	0.784	0.374
Extra-Trees + ling.	0.707	0.668	0.696	0.645	0.664	0.775	0.736	0.820	0.317
Stacking	0.689	0.638	0.693	0.594	0.660	0.789	0.726	0.864	0.272
Stacking + ling.	0.689	0.641	0.684	0.605	0.672	0.781	0.741	0.828	0.272
Emb. 1: *FacebookAI/roberta-large*									
SVM	0.665	0.631	0.632	0.634	**0.701**	0.788	0.771	0.806	0.436
SVM + ling.	0.677	0.645	0.644	0.649	0.700	0.775	**0.780**	0.770	0.455
MLP	0.655	0.610	0.655	0.594	0.681	**0.808**	0.748	**0.886**	0.282
MLP + ling.	0.656	0.609	0.654	0.585	0.697	0.800	0.761	0.842	0.291
Extra-Trees	0.641	0.583	0.631	0.544	0.660	0.784	0.726	0.852	0.273
Extra-Trees + ling.	0.647	0.590	0.637	0.553	0.660	0.783	0.724	0.852	0.276
Stacking	0.666	0.615	0.657	0.582	0.681	0.801	0.739	0.876	0.285
Stacking + ling.	0.670	0.626	0.653	0.605	0.672	0.799	0.731	**0.882**	0.294
Emb. 2: *cardiffnlp/twitter-roberta-base-hate-multiclass-latest*									
SVM	**0.728**	**0.705**	0.694	**0.719**	0.700	0.780	0.775	0.786	0.476
SVM + ling.	0.721	0.696	0.691	0.702	0.686	0.766	0.768	0.764	**0.490**
MLP	0.703	0.667	0.688	0.657	0.644	0.777	0.717	0.848	0.271
MLP + ling.	0.696	0.652	0.699	0.623	0.656	0.788	0.723	0.868	0.280
Extra-Trees	0.719	0.692	0.692	0.694	0.664	0.788	0.730	0.858	0.326
Extra-Trees + ling.	0.723	0.698	0.698	0.701	0.647	0.770	0.722	0.826	0.318
Stacking	0.723	0.692	0.705	0.681	0.660	0.789	0.726	0.864	0.262
Stacking + ling.	0.720	0.684	**0.709**	0.663	0.654	0.783	0.725	0.854	0.269

Table 5.2: Results of the models across three tasks related to sexism on TikTok in the Spanish corpus. 'S' refers to Sexist and 'D' denotes Direct sexism, which are the positive classes. Values in bold highlight the top-performing model for each metric across the tasks, while underlined values indicate the second highest performance.

Model	Task 1				Task 2				Task 3
	M-F1	F1 (S)	P (S)	R (S)	M-F1	F1 (D)	P (D)	R (D)	M-F1
TF-IDF									
SVM	0.646	0.704	0.708	0.701	0.670	0.810	0.781	0.843	0.293
SVM + ling.	0.656	0.700	0.728	0.674	0.692	0.823	0.794	0.856	0.300
MLP	0.644	0.718	0.698	0.740	0.661	0.813	0.773	0.859	0.253
MLP + ling.	0.658	0.725	0.711	0.740	0.676	0.821	0.781	0.867	0.266
Extra-Trees	0.689	0.723	0.765	0.687	**0.696**	0.816	0.803	0.830	0.395
Extra-Trees + ling.	0.690	0.728	0.762	0.698	0.689	0.820	0.793	0.851	0.379
Stacking	0.673	0.740	0.720	0.761	0.683	0.827	0.782	0.877	0.267
Stacking + ling.	0.686	<u>0.750</u>	0.730	0.771	0.687	0.826	0.786	0.871	0.269
Emb. 1: *PlanTL-GOB-ES/roberta-large-bne*									
SVM	0.693	0.718	<u>0.783</u>	0.665	0.694	0.783	**0.838**	0.737	<u>0.483</u>
SVM + ling.	0.691	0.712	**0.788**	0.652	0.683	0.773	0.832	0.724	**0.498**
MLP	0.669	0.697	0.768	0.650	0.640	**0.830**	0.761	**0.918**	0.275
MLP + ling.	0.672	0.731	0.734	0.738	0.651	0.819	0.767	0.881	0.305
Extra-Trees	0.6271	0.732	0.676	<u>0.800</u>	0.663	0.824	0.770	0.887	0.262
Extra-Trees + ling.	0.628	0.735	0.675	**0.808**	0.661	0.828	0.766	0.900	0.262
Stacking	0.671	0.738	0.721	0.759	0.662	0.826	0.768	0.894	0.285
Stacking + ling.	0.664	0.737	0.713	0.764	0.658	0.820	0.767	0.882	0.296
Emb. 2: *piuba-bigdata/beto-contextualized-hate-speech*									
SVM	0.685	0.708	0.781	0.648	0.690	0.795	0.817	0.777	0.448
SVM + ling.	<u>0.694</u>	0.720	0.784	0.668	<u>0.694</u>	0.785	<u>0.833</u>	0.743	0.468
MLP	0.681	0.717	0.768	0.682	0.650	0.811	0.770	0.862	0.285
MLP + ling.	0.682	0.740	0.744	0.743	0.632	0.807	0.759	0.868	0.278
Extra-Trees	0.673	0.736	0.724	0.748	0.648	0.819	0.763	0.886	0.269
Extra-Trees + ling.	0.679	0.743	0.727	0.760	0.653	0.824	0.763	0.897	0.277
Stacking	0.687	0.743	0.741	0.747	0.659	<u>0.829</u>	0.766	0.904	0.280
Stacking + ling.	**0.696**	**0.757**	0.742	0.773	0.648	0.826	0.760	<u>0.906</u>	0.286

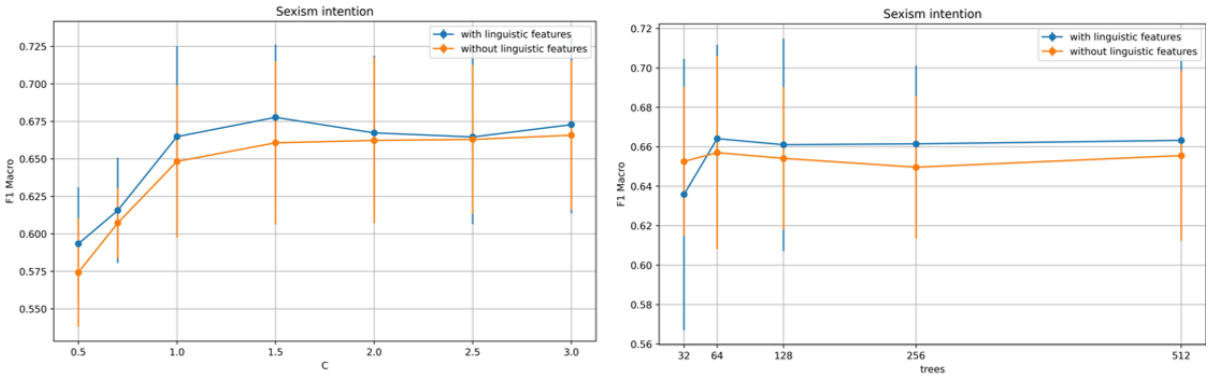


Figure 5.1: Grid Search Cross Validation Results with TF-IDF on English TikToks for SVM (Left) and ExtraTrees (Right)

5.2.1 Feature Importance and Partial Dependence Plots

Though TF-IDF representation might not achieve the same level of effectiveness as embeddings, examining its application remains worthwhile. In this instance, the Extra-Trees algorithm is utilized alongside linguistic variables.

5.2.1.1 Feature Importance

Figure 5.2 displays the impurity-based feature importances generated by Extra-Trees for Task 1 and Task 2. Feature importance is a measure of how much each feature contributes to the decision made by the algorithm. It is calculated based on the decrease in impurity caused by splitting on a particular feature. Higher feature importance indicates that the feature is more influential in making predictions.

For Task 1 (sexism detection), the most important features in Spanish include the hate speech score towards women, the words "las mujeres" (women), "hombres" (men), "feminista" (feminist), etc. In English, important features comprise the words "men", "a woman", sentiment score of joy, usage of the pronoun "I" etc.

For Task 2 (detection of sexism intention), in Spanish, crucial features encompass "pibas" (girls), "abuso" (abuse), usage of the purple heart emoji, etc. In English, significant features consist of the word "girlfriend", "vs", "care about", usage of present tense and adverbs, etc.

These insights shed light on the specific linguistic cues and contextual factors that the model relies on to make predictions related to sexism detection in both Spanish and English.

5.2.1.2 Partial Dependence Plots

Figure 5.3 illustrates the Partial Dependence Plots (PDPs) generated by Extra-Trees for Task 1 and Task 2. PDPs show the marginal effect of a feature on the predicted outcome while marginalizing out the effects of all other features. In other words, PDPs provide insights into how the predicted outcome changes as a single feature varies while keeping all other features constant.

For Task 1 in Spanish, we observe that words such as "las mujeres", "machista", the emotion score for love, and words related to humans from LIWC (adulta**, cría*, hombre*, human*, individu*, infantil*, juvenil*, masculino*, muchach*, mujer*...) influence the model's behavior. We can see that if the words "las mujeres" or "machista" appear, the model assigns a higher probability to TikToks being sexist. Similarly, a higher score of love decreases the probability of a TikTok being sexist. In English, words like "women are", hate score towards women, and anger-related words from LIWC influence the model. Higher frequencies of these words increase the likelihood of a TikTok being sexist. Additionally, a higher score of joy decreases the probability of a TikTok being sexist.

For the second task of detecting the intention of sexism in Spanish, we observe that more frequent use of words like "abuso" or "pibas", or the use of feeling-related words from LIWC (abraz*, acaric*, agarrába*, frotar, tocaré*, cogiera*...) lead to a higher probability of a TikTok being direct sexist. On the other hand, using the purple heart emoji or the word "abuso" increases the probability of the TikTok being reported sexist. In English, using words like "babe" or more frequent use of the pronoun "I" increases the likelihood of a TikTok being directly sexist, while increased usage of adverbs or the phrase "care about" increases the likelihood of it being reported sexist.

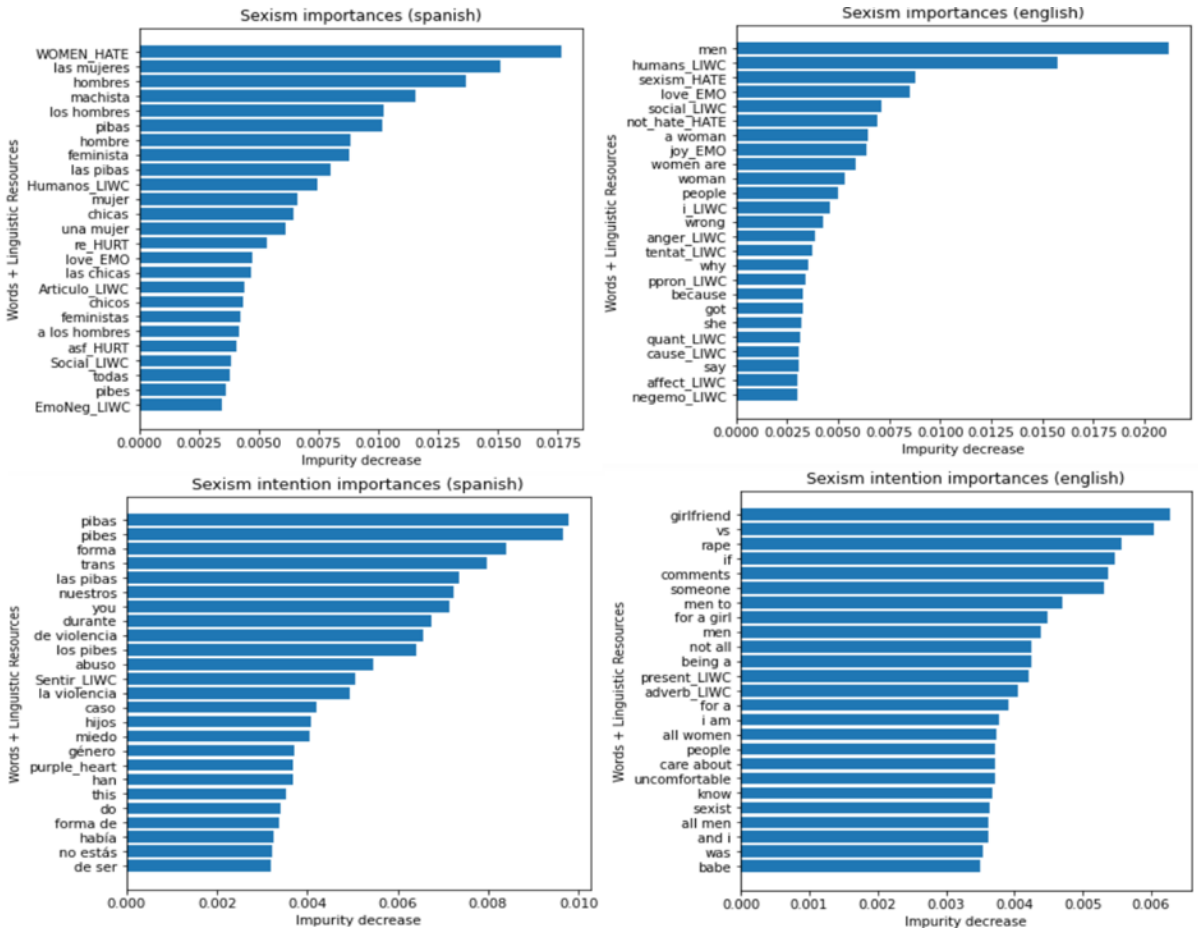


Figure 5.2: Extra-trees impurity-based feature importances for Task 1 and Task 2

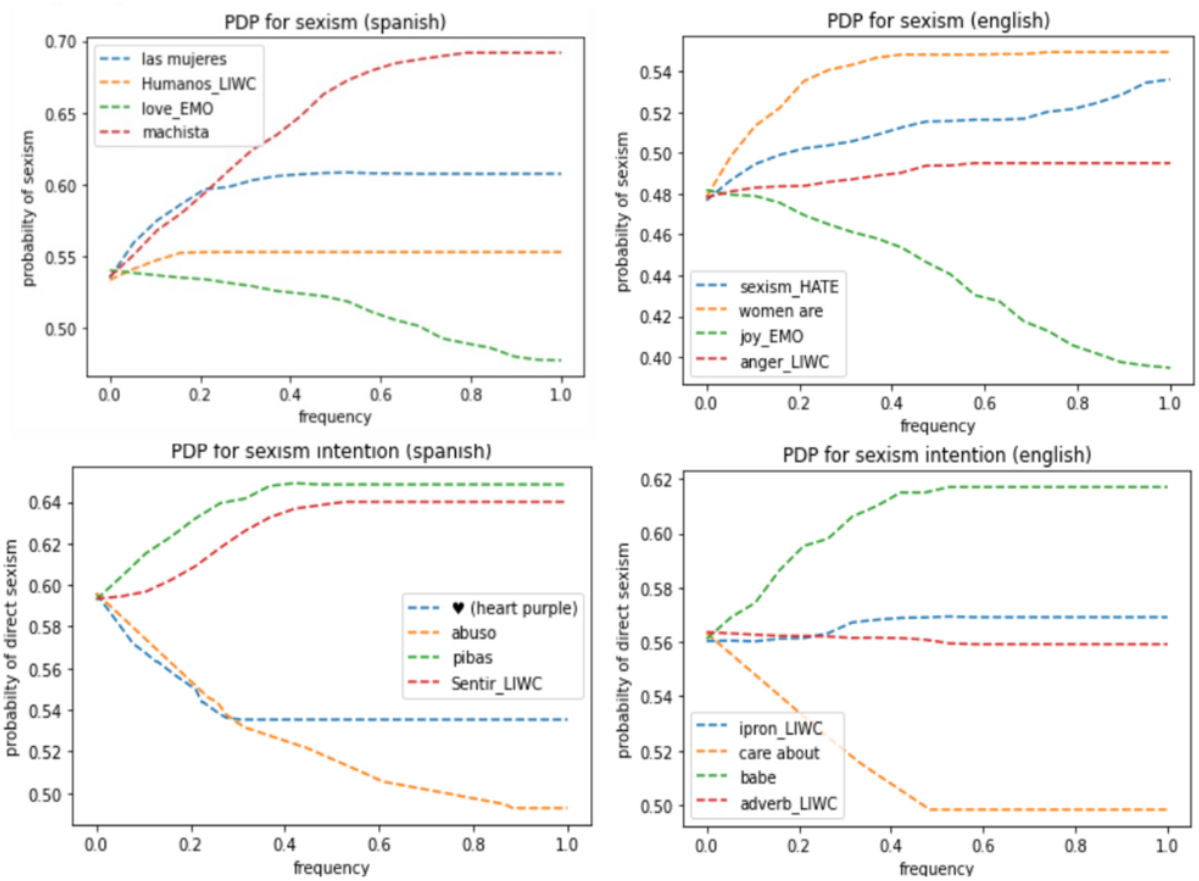


Figure 5.3: Extra-trees PDPs for Task 1 and Task 2

5.3 Audio Results

This section presents the results using Mel-Frequency Cepstral Coefficients and Pre-trained Wav2Vec2 Embeddings for the three sexism tasks. Refer to Table 5.3 for results in English and Table 5.4 for results in Spanish.

Table 5.3: Results of the models across three tasks related to sexism on TikTok in the English corpus. 'S' refers to Sexist and 'D' denotes Direct sexism, which are the positive classes. Values in bold highlight the top-performing model for each metric across the tasks, while underlined values indicate the second highest performance.

Model	Task 1				Task 2				Task 3
	M-F1	F1 (S)	P (S)	R (S)	M-F1	F1 (D)	P (D)	R (D)	M-F1
MFCCs									
SVM	0.572	**0.571**	0.523	**0.630**	0.580	0.689	0.687	0.693	0.364
MLP	0.565	0.532	0.535	0.543	0.529	0.743	0.660	0.856	0.182
Extra-trees	0.590	0.521	0.571	0.483	0.564	0.736	0.671	0.816	0.318
Stacking	0.570	0.460	0.590	0.380	0.468	0.768	0.642	**0.954**	0.192
Wav2Vec2									
SVM	**0.608**	0.567	0.572	0.565	**0.654**	0.733	**0.750**	0.719	**0.381**
MLP	0.585	0.536	0.575	0.544	0.603	0.752	0.694	0.827	0.237
Extra-trees	0.602	0.530	0.593	0.480	0.616	0.768	0.699	0.852	0.310
Stacking	0.599	0.522	**0.598**	0.470	0.605	**0.777**	0.690	0.890	0.261

Table 5.4: Results of the models across three tasks related to sexism on TikTok in the Spanish corpus. 'S' refers to Sexist and 'D' denotes Direct sexism, which are the positive classes. Values in bold highlight the top-performing model for each metric across the tasks, while underlined values indicate the second highest performance.

Model	Task 1				Task 2				Task 3
	M-F1	F1 (S)	P (S)	R (S)	M-F1	F1 (D)	P (D)	R (D)	M-F1
MFCCs									
SVM	**0.579**	0.648	**0.653**	0.645	0.643	0.774	0.776	0.773	0.340
MLP	0.513	0.674	0.616	0.762	0.621	0.816	0.754	0.895	0.153
Extra-trees	0.546	0.691	0.621	0.780	0.645	0.823	0.759	0.900	0.285
Stacking	0.508	**0.715**	0.608	**0.867**	0.628	0.825	0.748	**0.922**	0.182
Wav2Vec2									
SVM	0.572	0.648	0.645	0.652	**0.687**	0.810	**0.800**	0.820	**0.367**
MLP	0.526	0.692	0.621	0.793	0.669	0.821	0.777	0.872	0.207
Extra-trees	0.540	0.686	0.618	0.773	0.656	0.832	0.765	0.913	0.245
Stacking	0.516	0.700	0.610	0.823	0.675	**0.837**	0.773	0.914	0.215

The results for sexism tasks using audio variables exhibit interesting patterns across the different models and tasks. For the sexism detection task, the audio variables, both MFCCs and Wav2Vec2 embeddings, do not perform as strongly, achieving an F1-macro around 0.6. This suggests that audio features might not capture the nuanced patterns of sexism as effectively as text-based features for this particular task.

However, when considering the task of detecting the intention behind sexism, the audio-based models present a more competitive performance. Notably, the Wav2Vec2 embeddings outperform the MFCCs significantly. In English, the SVM model using Wav2Vec2 achieves the best performance with an F1-macro of 0.654, while in Spanish, it reaches 0.687, also using SVM. This demonstrates that audio embeddings like Wav2Vec2 can approximate the performance of text-based models for this specific task.

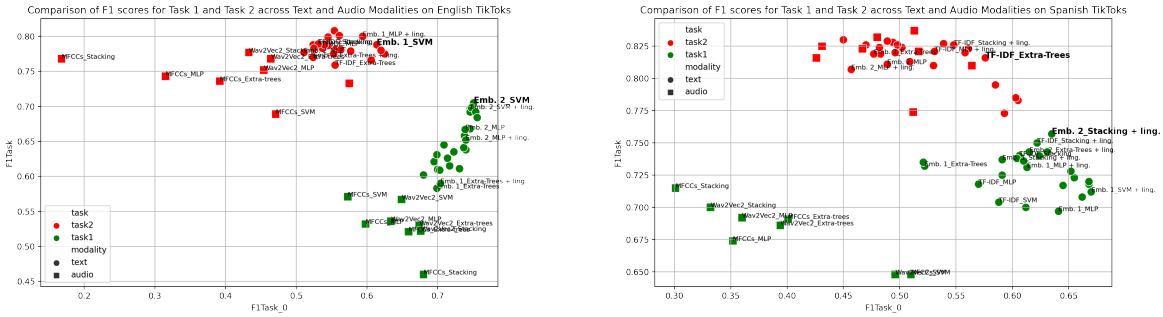
The graphical representation of these results in Figure 5.4 further elucidates these observations. In the English dataset, the audio models consistently underperform compared to their text-based counterparts across both tasks. This is expected, as text data can capture more explicit and diverse linguistic cues related to sexism compared to audio features.

On the other hand, the Spanish dataset presents a more intriguing scenario. With Wav2Vec2 embed-

dings, the audio models perform remarkably well, especially for the second task, achieving results that are comparable to those obtained with embeddings and TF-IDF applied to title+transcription+OCR. This is surprising because Wav2Vec2 is primarily an audio-based embedding model, and its ability to rival text-based methods suggests that it can capture significant linguistic and contextual information from audio data.

Overall, while audio features may not be as effective as text-based features for sexism detection, they show promise in capturing the subtleties and intentions behind sexist content, particularly when leveraging advanced audio embeddings like Wav2Vec2.

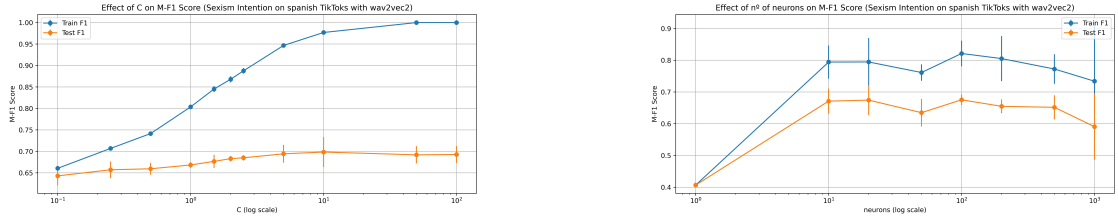
The grid search results for the Wav2Vec2 model applied to the Spanish Sexism Intention Task are shown in Figure 5.5. For the SVM model, increasing the regularization parameter C leads to a better fit to the training data, but the F1-macro score plateaus after $C = 10$, so this value is chosen for regularization. In the MLP model, using more than 200 neurons in the hidden layer does not improve the F1-macro score, while in Extra-Trees, increasing the number of estimators beyond 500 does not yield further improvements. These findings guide the selection of optimal hyperparameter values for the Wav2Vec2 model across SVM, MLP, and Extra-Trees, enhancing the model's effectiveness in the Spanish Sexism Intention Task.



(a) English TikToks

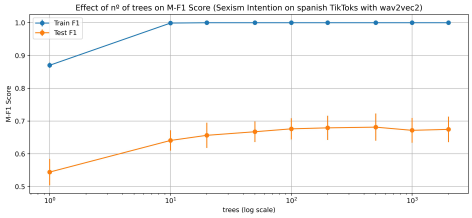
(b) Spanish TikToks

Figure 5.4: F1 Scores for Negative and Positive Classes (text and audio models)



(a) Grid Search Results for SVM

(b) Grid Search Results for MLP



(c) Grid Search Results for Extra-Trees

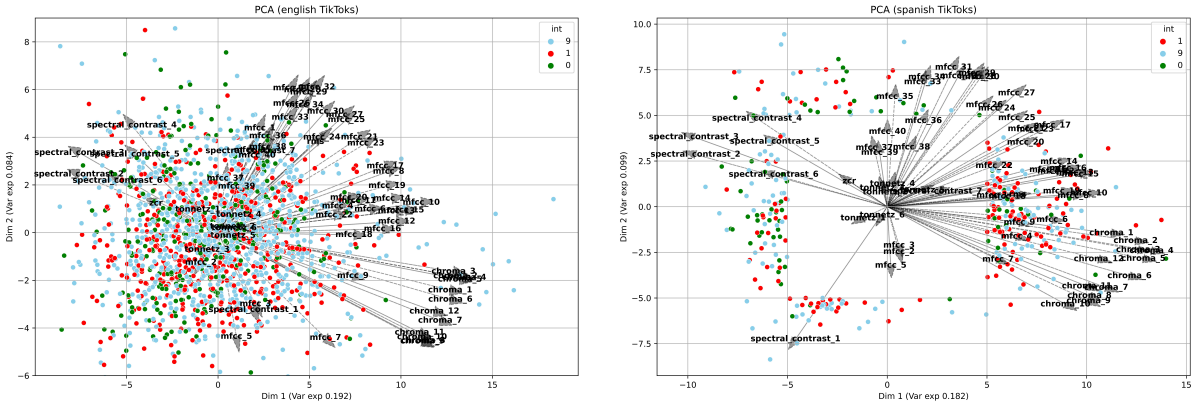
Figure 5.5: Grid Search Results for Wav2Vec2 in Spanish Sexism Intention Task

5.3.1 PCA and correlations

In Figures 5.6(a) and 5.6(b), we visualize the first two principal components, representing different aspects of the audio features. The first component explains roughly 20% of the variance, while the second component explains about 10%. These figures provide insights into which features contribute most to each component and reveal the correlation structure among the audio variables.

It appears that features associated with spectral contrast display strong positive correlations among themselves, while they are inversely correlated with chroma features. This indicates that spectral contrast features capture variations in the spectral texture or brightness of the audio signal, while chroma features represent the tonal content or musical notes present in the audio.

In the PCA plot of Spanish TikToks, for example, we observe that videos addressing sexism tend to have higher values of spectral contrast, particularly on the left side of the plot. This suggests that TikTok videos denouncing sexism exhibit more pronounced variations in spectral texture, possibly indicating higher emotional intensity or stronger language cues related to sexism. Conversely, videos with lower values of spectral contrast may represent content with less explicit or nuanced expressions of sexism.



(a) English TikToks

(b) Spanish TikToks

Figure 5.6: PCA on MFCCs features

In Figures 5.7 and 5.8, correlation heatmaps are displayed, illustrating the relationships between audio features and linguistic resources extracted from text for English and Spanish TikTok videos, respectively.

These correlations indicate the strength and direction of relationships between audio features and linguistic resources. It is important to note that the observed correlations are relatively weak, typically ranging between -0.25 and 0.25, suggesting that there is no strong linear relationship between audio features and linguistic resources.

Of particular interest is the observation that features related to spectral contrast, previously seen in PCA analysis to be associated with reported instances of sexism, exhibit correlations with linguistic categories from LIWC (Linguistic Inquiry and Word Count) analyses. For example, in the case of Spanish TikToks, spectral contrast features 1, 2, and 3 show positive correlations with the usage of words related to physical contact and personal pronouns. This suggests that TikTok videos with higher spectral contrast values may contain more explicit language or discussions related to personal experiences, possibly indicating the presence of more emotionally charged or confrontational content regarding sexism.

Conversely, chroma features show inverse correlations with these linguistic categories. For instance, in Spanish TikToks, chroma features exhibit negative correlations with words related to physical contact (MecCog) and personal pronouns (PronImp). This could imply that TikTok videos with more musical content, as indicated by higher chroma values, tend to have less emphasis on explicit language or personal experiences regarding sexism. Instead, they may prioritize entertainment or artistic expression through music and visuals, possibly presenting sexism in a more implicit or subtle manner.

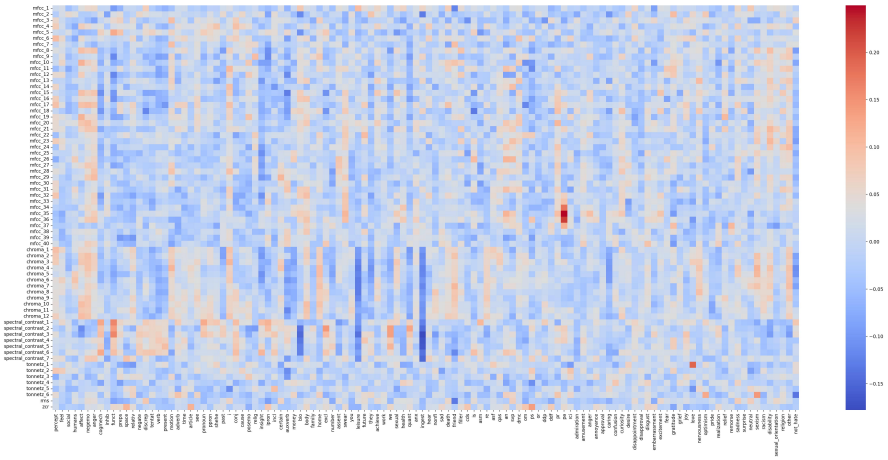


Figure 5.7: Relationship between audio features and linguistic resources features extracted from text for English.

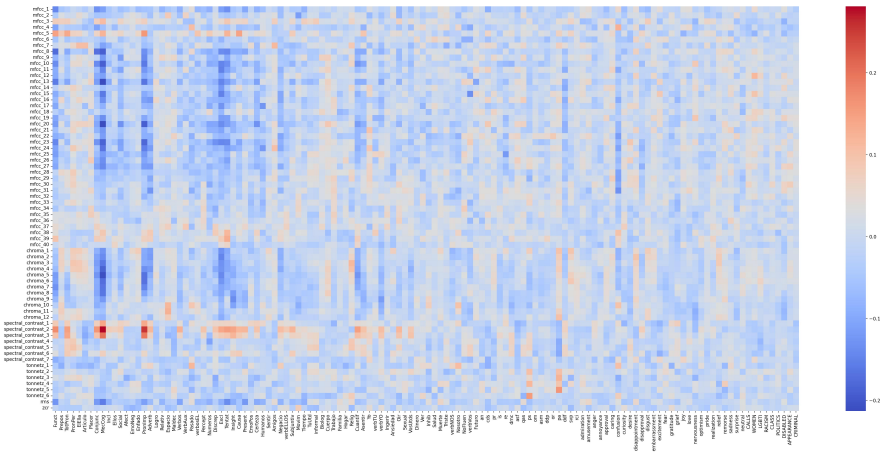


Figure 5.8: Relationship between audio features and linguistic resources features extracted from text for Spanish.

5.3.2 Feature Importance and Partial Dependence Plots

In Figure 5.9, the Permutation Importances for the SVM model are presented. The importance scores for the MFCCs features were calculated using 10 repetitions of permutation, allowing us to assess the importance of each feature for the model's performance. It is worth noting that many of these variables are correlated with each other, leading to increased variability when calculating these importances.

On the other hand, in Figure 5.10, the Partial Dependence Plots for the audio features are shown. It's interesting to note that both in English and Spanish, when determining whether sexism is direct or reported, a similar pattern is observed. As the RMS increases, the model tends to assign a higher probability to the TikTok being sexist. This could be because many TikToks containing direct sexism are jokes that manipulate audio power, including loud screams or music. Conversely, as the Zero Crossing Rate or spectral contrast increases, it is more likely that a TikTok will be classified as reported sexism, suggesting that these features are related to TikToks denouncing sexism. Higher ZCR values typically indicate a higher frequency of rapid changes in the audio waveform, which can be indicative of speech or vocal activity. In the context of TikTok videos denouncing sexism, creators may employ speech to report instances of sexism, share personal experiences, or advocate for change. Consequently, TikToks with reported sexism are likely to have higher ZCR values due to the presence of speech segments denouncing

sexism. Therefore, the SVM model may learn to associate higher ZCR values with a higher likelihood of containing reported instances of sexism.

The distributions depicted in Figures 5.11, 5.12, and 5.13 reveal significant differences in audio features across various types of video content. In Figure 5.11, non-sexist videos tend to exhibit higher frequencies in spectral contrasts 4, 5, and 7. This suggests that non-sexist content may possess a distinct spectral signature. Moving to Figure 5.12, directly sexist videos show increased values in MFCC19 and MFCC22, as well as chroma features such as Chroma5, Chroma7, and Chroma10. Reported sexist videos, on the other hand, present higher levels in spectral contrasts 1 through 6. This differentiation highlights the nuanced audio patterns that distinguish directly sexist content, which also shows a significantly higher RMS energy with a p-value of 0.0005. However, no significant difference was observed in the ZCR between directly and reported sexist content. Lastly, Figure 5.13 examines sexism categories, for example videos with Sexual Violence tend to show a higher frequency of MFCC11 and lower than MFCC19, alongside lower ZCR values, underscoring the specific audio characteristics associated with this severe category of content.

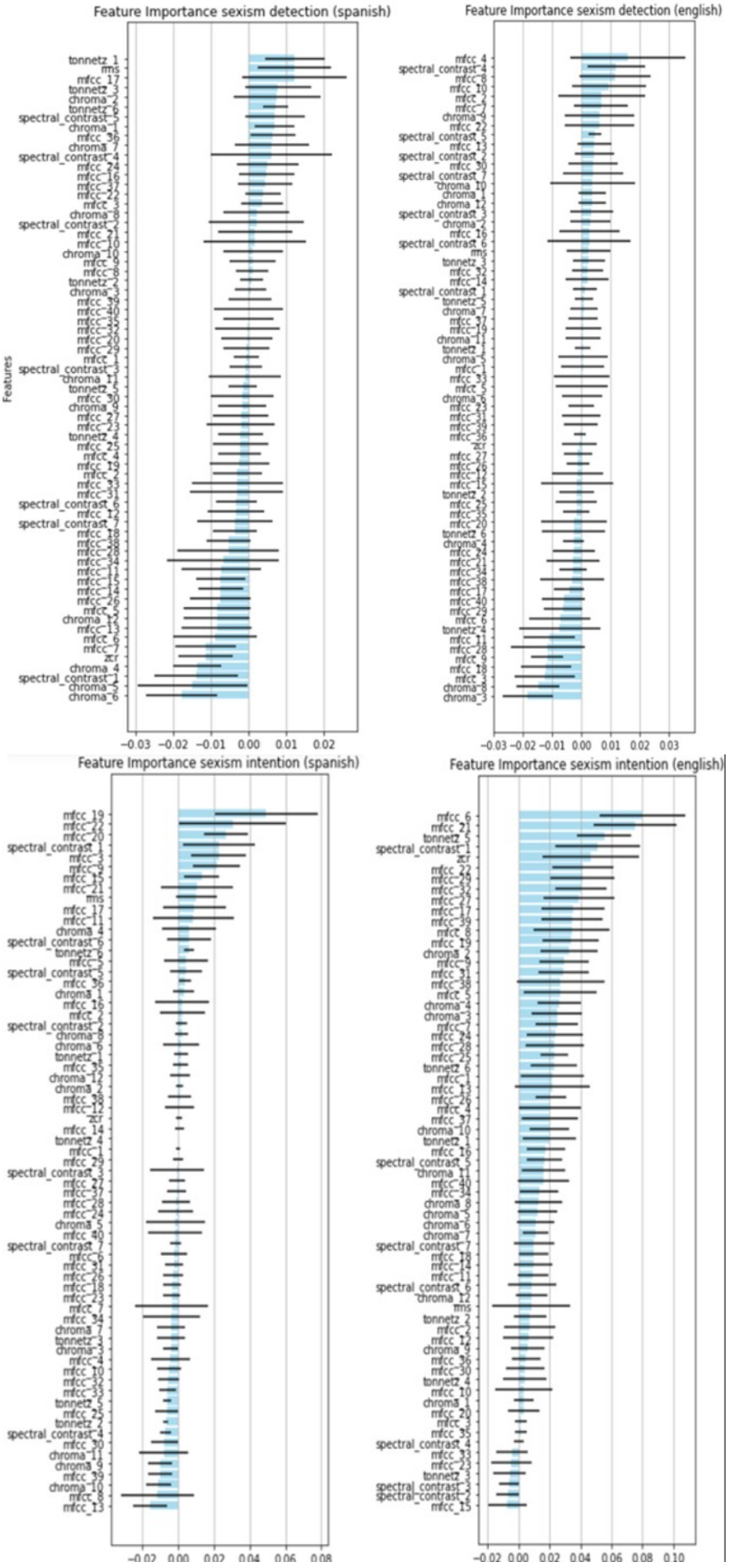


Figure 5.9: Permutation Importances for SVM Model

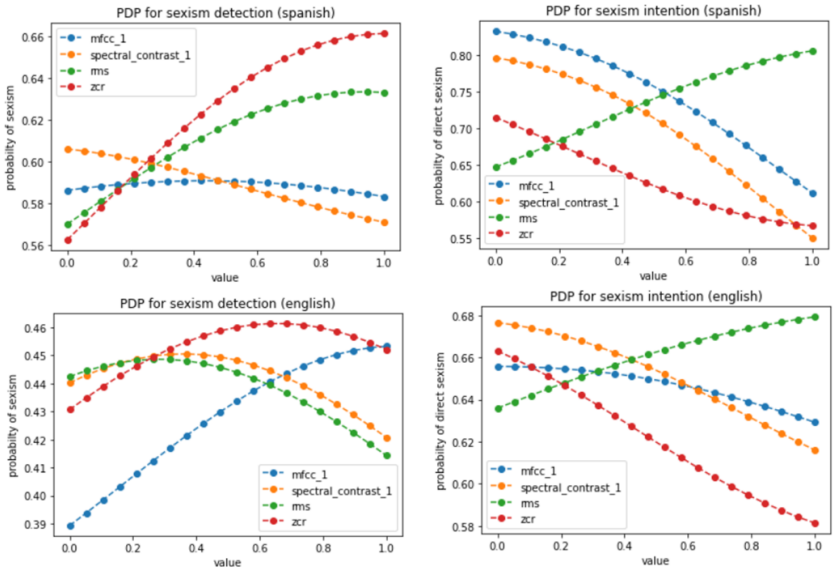


Figure 5.10: Partial Dependence Plots for Audio Features

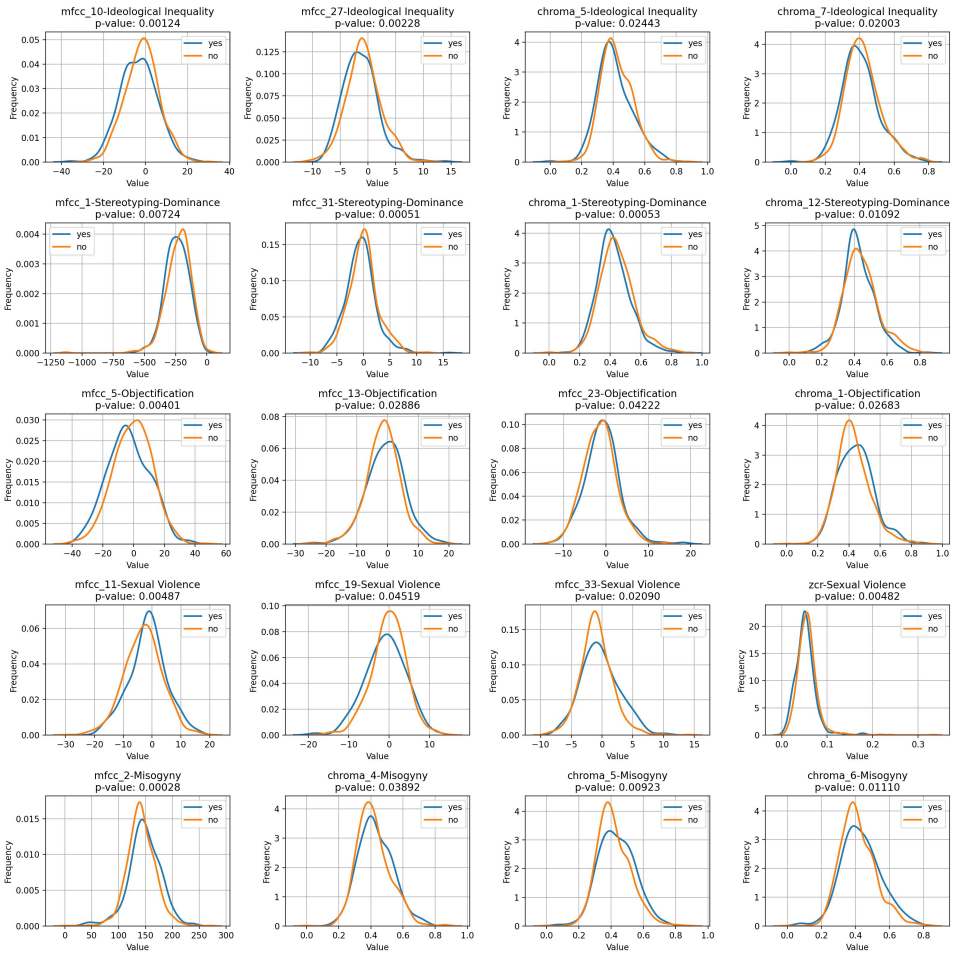


Figure 5.11: Distribution of MFCCs for Task 1

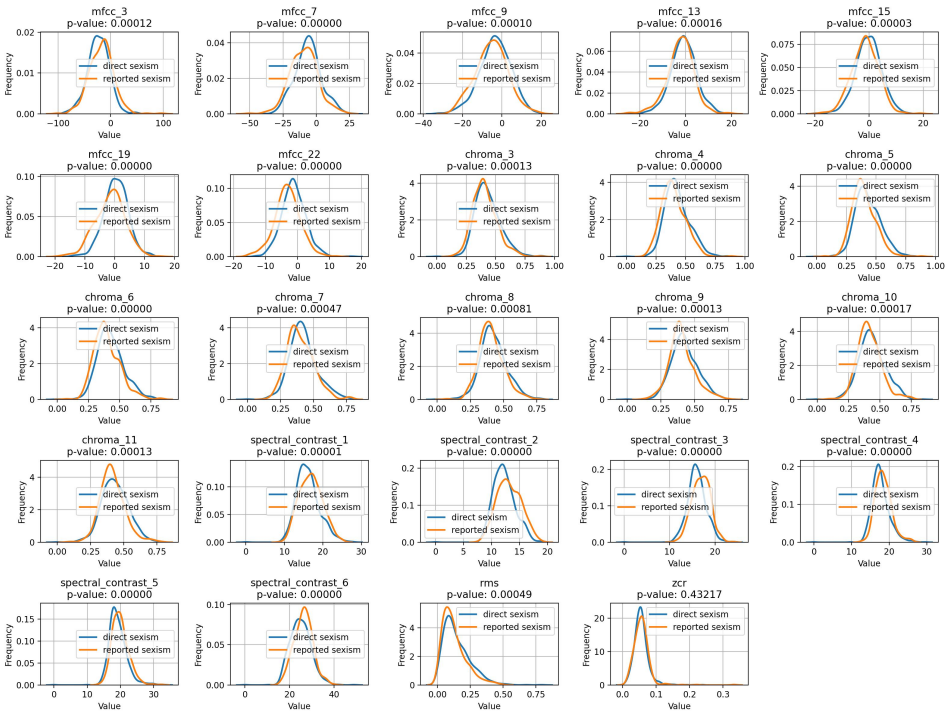


Figure 5.12: Distribution of MFCCs for Task 2

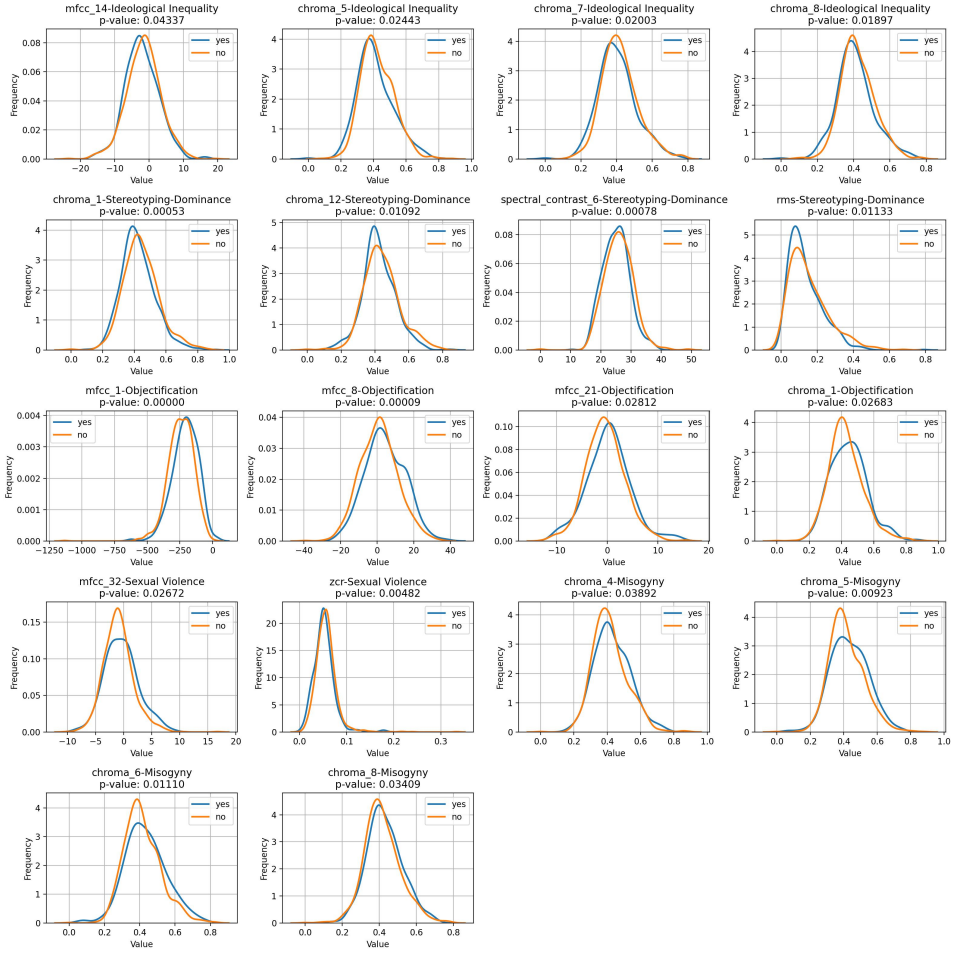


Figure 5.13: Distribution of MFCCs for Task 3

5.4 Video Results

In this section, the findings from applying the three distinct video models discussed earlier to the task of identifying instances of sexism within TikTok content will be explored. Each model employs a unique approach to analyzing video data, utilizing different methodologies and architectures. The results presented here offer insights into the effectiveness of these models in detecting sexism and understanding the nuanced aspects of content on the platform. The results for the three tasks of sexism detection, both in English and Spanish, will be explored. For the results in English, refer to Table 5.5. For the results in Spanish, refer to Table 5.6.

Table 5.5: Results of the video models across three tasks related to sexism on TikTok in the English corpus. 'S' refers to Sexist and 'D' denotes Direct sexism, which are the positive classes. Values in bold highlight the top-performing model for each metric across the tasks, while underlined values indicate the second highest performance.

Model	Task 1				Task 2				Task 3
	M-F1	F1 (S)	P (S)	R (S)	M-F1	F1 (D)	P (D)	R (D)	M-F1
ResNet+LSTM									
ResNet+LSTM	0.555	0.543	0.526	**0.574**	0.622	0.651	0.749	0.660	0.184
ViT+LSTM	<u>0.572</u>	<u>0.550</u>	<u>0.540</u>	0.541	<u>0.668</u>	<u>0.707</u>	**0.765**	<u>0.743</u>	<u>0.269</u>
BLIP+TF-IDF	**0.630**	**0.578**	**0.611**	<u>0.553</u>	**0.687**	**0.794**	<u>0.749</u>	**0.846**	**0.302**

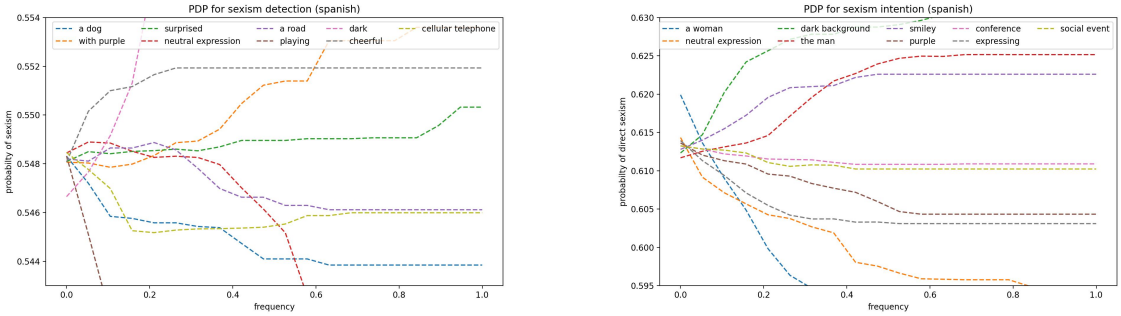
Table 5.6: Results of the video models across three tasks related to sexism on TikTok in the Spanish corpus. 'S' refers to Sexist and 'D' denotes Direct sexism, which are the positive classes. Values in bold highlight the top-performing model for each metric across the tasks, while underlined values indicate the second highest performance.

Model	Task 1				Task 2				Task 3
	M-F1	F1 (S)	P (S)	R (S)	M-F1	F1 (D)	P (D)	R (D)	M-F1
ResNet+LSTM									
ResNet+LSTM	0.551	0.601	<u>0.646</u>	0.681	0.625	0.697	0.789	0.786	<u>0.217</u>
ViT+LSTM	<u>0.551</u>	<u>0.603</u>	0.628	<u>0.696</u>	<u>0.678</u>	<u>0.745</u>	<u>0.790</u>	<u>0.837</u>	**0.312**
BLIP+TF-IDF	**0.592**	**0.708**	**0.650**	**0.778**	**0.693**	**0.827**	**0.791**	**0.867**	0.209

For the English dataset, the best-performing model for both Task 1 and Task 2 was found to be BLIP+TF-IDF, achieving macro F1 scores of 0.630 and 0.687, respectively. Similarly, in the Spanish dataset, BLIP+TF-IDF emerged as the top-performing model for both Task 1 and Task 2, with macro F1 scores of 0.592 and 0.693, respectively. However, for Task 3 in English, BLIP+TF-IDF obtained the highest F1-macro score of 0.302, while in Spanish, ViT+LSTM achieved the highest F1-macro score of 0.312.

Figure 5.14 shows that in Spanish TikTok videos, terms like "dark", "purple", and "cheerful" in video captions are associated with sexist classifications by the BLIP model, while neutral terms such as "a dog" or "road" are linked to non-sexist content. Additionally, "dark background" and "smiley" relate to direct sexism, whereas "a woman" and "conference" are indicative of reported sexism. Similar patterns are observed in English TikToks.

In Figures 5.15(a) and 5.15(b), we observe the distribution of terms detected by the BLIP model across the sexism detection and intention tasks. For instance, the term "a dog" identified by the BLIP model rarely appears in sexist videos; however, its distribution in non-sexist videos shows varying frequencies greater than zero, with a significant p-value of 0.036. This suggests a higher likelihood of the term appearing in non-sexist contexts. Similarly, when the term "conference" is detected, it is more likely to be associated with reported sexism, where the sexism is being denounced rather than being direct sexism, with a p-value of 0.012. These distributions illustrate how specific terms correlate differently with sexist and non-sexist content, providing insights into the contextual usage of language in these categorizations.



(a) Partial Dependence Plots (PDPs) for detected words by the BLIP model on sexism detection task.

(b) Partial Dependence Plots (PDPs) for detected words by the BLIP model on sexism intent task.

Figure 5.14: Partial Dependence Plots (PDPs) for BLIP model

5.5 Multimodal Results

5.5.1 Multimodal SVM

In our initial approach, we combined all features from text, audio, video, and linguistic resources as inputs for a Support Vector Machine classifier. For these results, we followed a 10-fold cross-validation procedure. This method aimed to comprehensively evaluate the impact of each feature set when used together, ensuring that the model's performance was robust and not dependent on a single data split. The combination of these diverse features provided a holistic view and enhanced the detection capabilities of the model across different modalities.

In Task 1, using only the text modality for detecting sexism presence yielded similar results to multimodal approaches, indicating that textual linguistic features were more crucial than audio or video cues. Similarly, for Task 3, which aimed to identify specific categories of sexism, text alone achieved similar performance (therefore, we do not show the results of the ablation tests here).

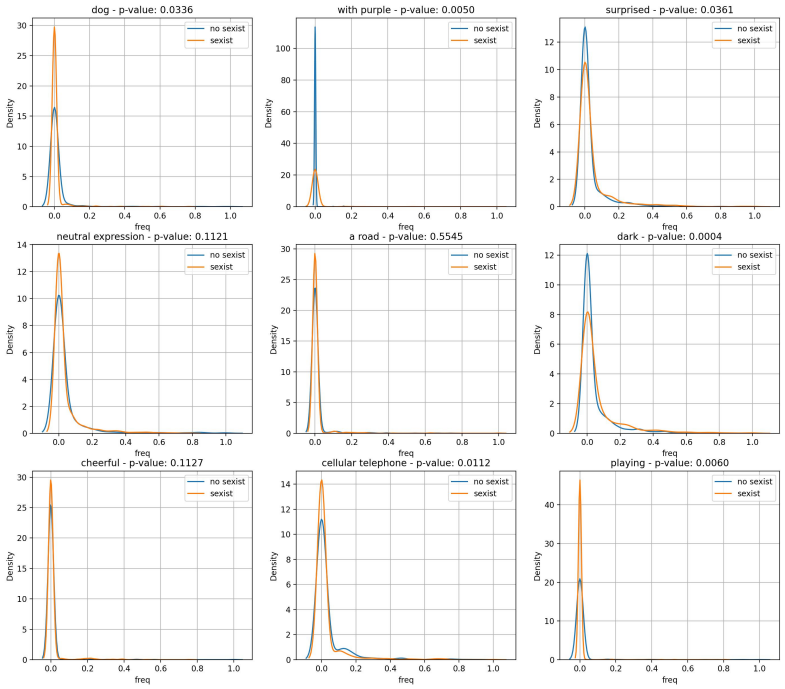
For Task 2 of detecting the intent of sexism, the multimodal approach significantly improved the outcomes compared to unimodal models for both English and Spanish languages. Figure 5.16 illustrates these findings, demonstrating the efficacy of integrating multiple modalities. The best-performing unimodal model was using the ViT in both English and Spanish, achieving an F1-macro score of 0.709 and 0.720, respectively. However, the highest performance was observed with the TAV model (combining Text, Audio, and Video) that excluded linguistic features (L). Specifically, for audio, the model employed MFCCs, and for video, it used ViT features. This configuration led to F1-macro scores of 0.753 and 0.768 in English and Spanish respectively, marking an improvement of 4.4% and 4.8% over the best unimodal models. The multimodal approach demonstrates the benefits of integrating text, audio, and video to achieve a more comprehensive understanding of content, leading to more accurate detection of sexist intent.

5.5.1.1 ROC Curves

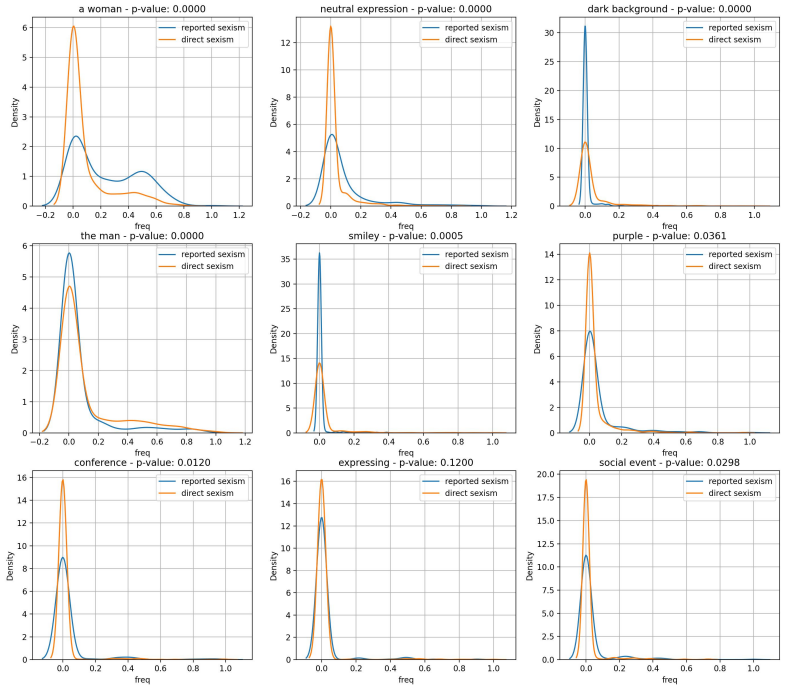
The Receiver Operating Characteristic (ROC) curves represent a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The Area Under the Curve (AUC) is a key metric that quantifies the overall ability of the classifier to discriminate between the classes. An AUC of 0.75, for example, indicates that in the task of detecting sexism, given one sexist and one non-sexist TikTok, the classifier will correctly assign a higher probability of being sexist to the actual sexist TikTok 75% of the time.

The ROC curves were calculated through cross-validation, ensuring that the probabilities were obtained for all instances within our dataset, thus providing a robust evaluation of the model's performance. Additionally, the ROC curves illustrate optimal thresholds determined according to the Youden Index, a combined measure of sensitivity and specificity, calculated as follows:

$$J = \max(sensitivity + specificity - 1)$$



(a) Distribution of detected words by BLIP model on sexism detection task



(b) Distribution of detected words by BLIP model on sexism intention task

Figure 5.15: Distribution of detected words by BLIP model

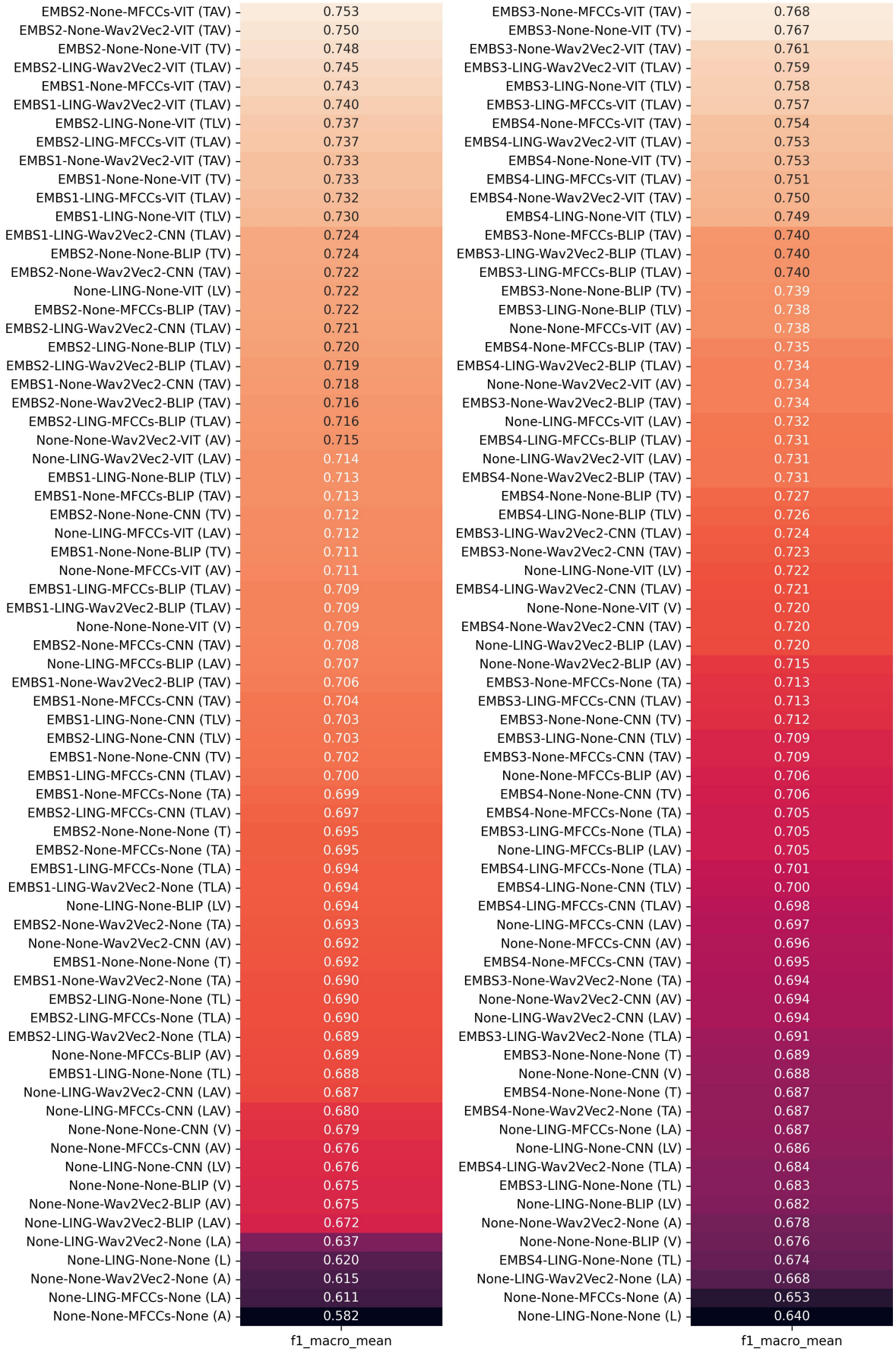


Figure 5.16: Multimodality Results for Task 2

where sensitivity is the true positive rate, and specificity is the true negative rate. Through cross-validation, various thresholds were tested to identify the one that maximizes the F1-Macro score, providing a balanced measure of precision and recall across all classes.

In the analysis of Task 1 (Figures 5.17(a) and 5.17(b)), the results are quite similar whether using only text embeddings or combining text, audio, and video, with AUCs nearing 0.8 in English and around 0.75 in Spanish. Notably, linguistic variables achieve better AUC values of about 0.67 in both languages compared to video and audio modalities, highlighting their effectiveness in this context.

Task 2 (Figures 5.17(c) and 5.17(d)) demonstrates that the best-performing unimodal models are those employing video with Vision Transformer (ViT), achieving AUCs of 0.768 in English and 0.782 in Spanish. Text modalities using embeddings follow closely, with AUCs of 0.753 in English and 0.772 in Spanish. Interestingly, linguistic variables outperform audio modalities using Wav2Vec2. The multimodal approach, which integrates text, audio, and video, significantly enhances the results, obtaining AUCs of 0.810 in English and 0.835 in Spanish, indicating a substantial improvement over the best unimodal results.

For Task 3, the figures (Figures 5.17(e), 5.17(f), 5.17(g), 5.17(h), and 5.17(i)) primarily focus on Spanish, with analogous findings for English. The use of multimodality does not significantly improve outcomes compared to using text embeddings alone in this preliminary approach. Ideological Inequality is the category best predicted across all modalities, with AUCs spanning from 0.69 for Wav2Vec2 to 0.83 for multimodal setups. Furthermore, the second-best predicted category using linguistic variables is Sexual Violence, with an AUC of 0.68, performing better than the other modalities.

This comprehensive analysis across different tasks and modalities highlights the varying effectiveness of each approach and the potential benefits of multimodal strategies in enhancing model performance.

5.5.2 Text, Audio, Video and Linguistic Features (TAVL) - Fine-Tuning

Now we present the results of our second approach, which involves fine-tuning some of the last attention layers of the three models (text, audio, and video) simultaneously. In this case, the results are shown only for a test set and not for cross-validation (due to computational resources and time constraints).

For training the model, we initially froze all parameters of the transformers and trained only the classification layers for a few epochs. Subsequently, we unfroze the last attention layers of each model, reduced the learning rate, and trained for a few more epochs until no improvement was observed on the validation set. It is worth noting that the batch size used was 4 (due to memory limitations), but with gradient accumulation. The loss function used is binary cross-entropy with weights to account for class imbalance, defined as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N} w_i \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right]$$

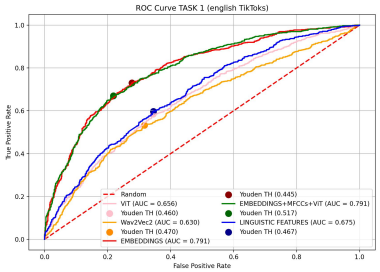
where w_i are the weights for class imbalance, y_i are the true labels, and p_i are the predicted probabilities.

We compare the best results obtained using only one modality, the approach of concatenating all pre-trained embeddings and training an SVM, and the fine-tuning method for English and Spanish across the three tasks (see Tables 5.7 and 5.8).

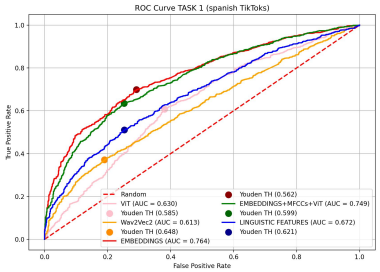
For Task 1, the fine-tuning multimodal approach improved over both the unimodal and SVM approaches. In English, using only text achieved an F1 macro score of 0.696, the SVM approach achieved 0.725, and the fine-tuning approach reached 0.751. Similarly, in Spanish, using only text achieved an F1 macro score of 0.700, while the SVM approach achieved 0.688, and the fine-tuning approach reached 0.722.

For Task 2, the results were similar. In English, the best unimodal model was using video, achieving an F1 macro score of 0.714. The SVM approach achieved 0.760, while the fine-tuning approach reached 0.787. In Spanish, the best unimodal model achieved 0.710, the SVM approach achieved 0.759, and the fine-tuning approach reached 0.804.

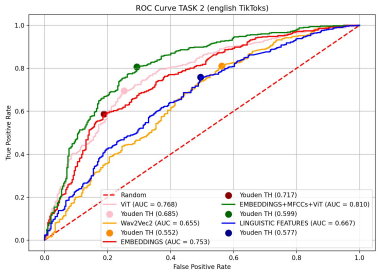
For Task 3 in English, the results were not as good, mainly due to the lower number of sexist videos compared to non-sexist videos in this language, and many categories being underrepresented. However, in Spanish, the results obtained, especially with fine-tuning, were acceptable and improved over the unimodal text-based approach and the SVM approach, achieving an F1 macro score around 0.700 for all categories.



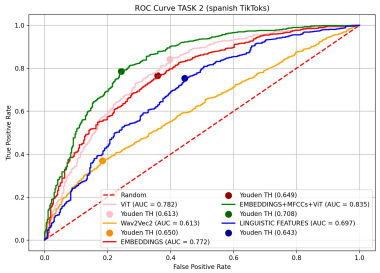
(a) ROC Curve for Task 1 (English)



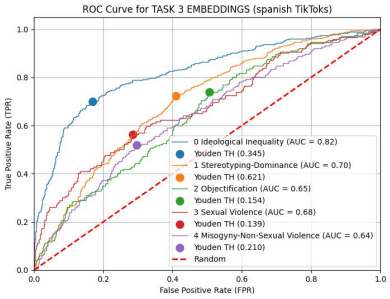
(b) ROC Curve for Task 1 (Spanish)



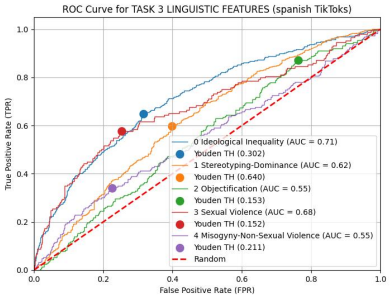
(c) ROC Curve for Task 2 (English)



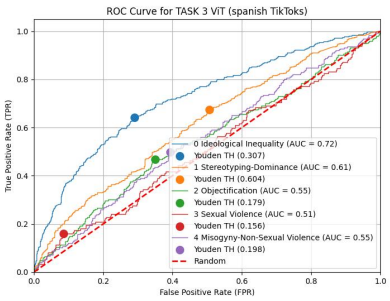
(d) ROC Curve for Task 2 (Spanish)



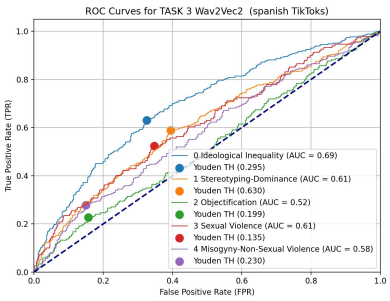
(e) ROC Curve for Task 3 with text embeddings (Spanish)



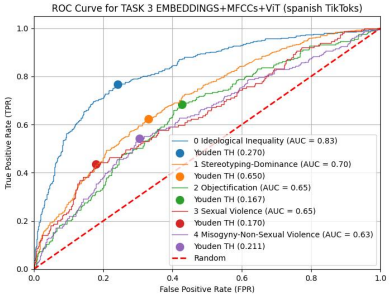
(f) ROC Curve for Task 3 with linguistic features (Spanish)



(g) ROC Curve for Task 3 with ViT (Spanish)



(h) ROC Curves for TASK 3 Wav2Vec2 (Spanish)



(i) ROC Curve for Task 3 with EMBS+MFCCs+ViT (Spanish)

Figure 5.17: ROC Curves across different tasks and languages

5.5.3 Confusion matrices

Analyzing the confusion matrices provides further insights into the performance of the models.

Task 1: Detecting Sexism Presence
In English (Table 5.7), the fine-tuning approach confusion matrix for Task 1 is:

$$\begin{bmatrix} 68 & 22 \\ 21 & 62 \end{bmatrix}$$

This indicates that the model correctly identified 68 instances of non-sexist content and 62 instances of sexist content, while misclassifying 22 instances of non-sexist content as sexist and 21 instances of sexist content as non-sexist.

Comparatively, the SVM approach had more misclassifications with a confusion matrix of:

$$\begin{bmatrix} 72 & 18 \\ 29 & 54 \end{bmatrix}$$

The unimodal text approach had a confusion matrix of:

$$\begin{bmatrix} 69 & 21 \\ 31 & 52 \end{bmatrix}$$

We can see that the fine-tuning approach significantly reduced the number of false negatives, which is crucial for sensitive tasks such as detecting sexism.

Task 2: Detecting the Intent of Sexism
In Spanish (Table 5.8), the fine-tuning approach confusion matrix for Task 2 is:

$$\begin{bmatrix} 25 & 11 \\ 8 & 72 \end{bmatrix}$$

This matrix shows an improvement in correctly identifying direct sexism compared to the SVM approach's confusion matrix of:

$$\begin{bmatrix} 22 & 14 \\ 9 & 71 \end{bmatrix}$$

The unimodal video approach had a confusion matrix of:

$$\begin{bmatrix} 22 & 14 \\ 15 & 65 \end{bmatrix}$$

The fine-tuning approach reduced both false positives and false negatives, enhancing the model's reliability in real-world applications.

Task 3: Identifying Specific Categories of Sexism
For Task 3, examining the results for the "Sexual Violence" category in Spanish (Table 5.8), the fine-tuning approach confusion matrix is:

$$\begin{bmatrix} 98 & 6 \\ 4 & 8 \end{bmatrix}$$

The SVM approach had a confusion matrix of:

$$\begin{bmatrix} 100 & 4 \\ 6 & 6 \end{bmatrix}$$

The unimodal text approach had a confusion matrix of:

$$\begin{bmatrix} 94 & 10 \\ 4 & 8 \end{bmatrix}$$

We can see that the fine-tuning approach provided a balance between true positives and reducing false positives, showing a better overall performance in identifying the specific category of sexual violence.

Table 5.7: Performance Results for Tasks in English. The bold values highlight the best-performing model for each task or category of sexism.

Task	Model	F1 Macro	Confusion Matrix
Task 1	SVM Multimodal (TAV)	0.725	$\begin{bmatrix} 72 & 18 \\ 29 & 54 \end{bmatrix}$
	TAVL - Fine-Tuning	**0.751**	$\begin{bmatrix} 68 & 22 \\ 21 & 62 \end{bmatrix}$
	Best Unimodal Model (T)	0.696	$\begin{bmatrix} 69 & 21 \\ 31 & 52 \end{bmatrix}$
Task 2	SVM Multimodal (TAV)	0.760	$\begin{bmatrix} 21 & 9 \\ 9 & 41 \end{bmatrix}$
	TAVL - Fine-Tuning	**0.787**	$\begin{bmatrix} 22 & 8 \\ 8 & 42 \end{bmatrix}$
	Best Unimodal Model (ViT)	0.714	$\begin{bmatrix} 18 & 12 \\ 9 & 41 \end{bmatrix}$
Task 3	SVM Multimodal (TAV)	0.499	Ideological Inequality
			$\begin{bmatrix} 27 & 16 \\ 23 & 14 \end{bmatrix}$
		0.558	Stereotyping-Dominance
			$\begin{bmatrix} 4 & 11 \\ 10 & 55 \end{bmatrix}$
		0.623	Objectification
			$\begin{bmatrix} 48 & 16 \\ 7 & 9 \end{bmatrix}$
		0.588	Sexual Violence
			$\begin{bmatrix} 72 & 2 \\ 5 & 1 \end{bmatrix}$
		0.492	Misogyny-Non-Sexual Violence
			$\begin{bmatrix} 62 & 6 \\ 11 & 1 \end{bmatrix}$
	TAVL - Fine-Tuning	0.487	Ideological Inequality
			$\begin{bmatrix} 18 & 25 \\ 16 & 21 \end{bmatrix}$
		0.536	Stereotyping-Dominance
			$\begin{bmatrix} 15 & 0 \\ 36 & 29 \end{bmatrix}$
		0.499	Objectification
			$\begin{bmatrix} 27 & 37 \\ 2 & 14 \end{bmatrix}$
		0.706	Sexual Violence
			$\begin{bmatrix} 73 & 1 \\ 4 & 2 \end{bmatrix}$
		0.424	Misogyny-Non-Sexual Violence
			$\begin{bmatrix} 59 & 9 \\ 12 & 0 \end{bmatrix}$
	Best Unimodal Model (T)	**0.540**	Ideological Inequality
			$\begin{bmatrix} 28 & 15 \\ 21 & 16 \end{bmatrix}$
		0.590	Stereotyping-Dominance
			$\begin{bmatrix} 8 & 7 \\ 18 & 47 \end{bmatrix}$
		0.675	Objectification
			$\begin{bmatrix} 47 & 17 \\ 4 & 12 \end{bmatrix}$
		0.530	Sexual Violence
			$\begin{bmatrix} 67 & 7 \\ 5 & 1 \end{bmatrix}$
		0.518	Misogyny-Non-Sexual Violence
			$\begin{bmatrix} 59 & 9 \\ 10 & 2 \end{bmatrix}$

Table 5.8: Performance Results for Tasks in Spanish. The bold values highlight the best-performing model for each task or category of sexism.

Task	Model	F1 Macro	Confusion Matrix
Task 1	SVM Multimodal (TAV)	0.688	$\begin{bmatrix} 56 & 31 \\ 35 & 97 \end{bmatrix}$
	TAVL - Fine-Tuning	**0.722**	$\begin{bmatrix} 57 & 30 \\ 28 & 104 \end{bmatrix}$
	Best Unimodal Model (T)	0.700	$\begin{bmatrix} 56 & 31 \\ 32 & 100 \end{bmatrix}$
Task 2	SVM Multimodal (TAV)	0.759	$\begin{bmatrix} 22 & 14 \\ 9 & 71 \end{bmatrix}$
	TAVL - Fine-Tuning	**0.804**	$\begin{bmatrix} 25 & 11 \\ 8 & 72 \end{bmatrix}$
	Best Unimodal Model (ViT)	0.710	$\begin{bmatrix} 22 & 14 \\ 15 & 65 \end{bmatrix}$
Task 3	SVM Multimodal (TAV)	0.787	Ideological Inequality
			$\begin{bmatrix} 70 & 5 \\ 16 & 25 \end{bmatrix}$
		0.701	Stereotyping-Dominance
			$\begin{bmatrix} 26 & 22 \\ 10 & 58 \end{bmatrix}$
		0.448	Objectification
			$\begin{bmatrix} 94 & 2 \\ 20 & 0 \end{bmatrix}$
		0.749	Sexual Violence
			$\begin{bmatrix} 100 & 4 \\ 6 & 6 \end{bmatrix}$
		0.506	Misogyny-Non-Sexual Violence
			$\begin{bmatrix} 78 & 7 \\ 27 & 4 \end{bmatrix}$
	TAVL - Fine-Tuning	0.742	Ideological Inequality
			$\begin{bmatrix} 69 & 6 \\ 19 & 22 \end{bmatrix}$
		0.708	Stereotyping-Dominance
			$\begin{bmatrix} 32 & 16 \\ 17 & 51 \end{bmatrix}$
		0.702	Objectification
			$\begin{bmatrix} 82 & 14 \\ 8 & 12 \end{bmatrix}$
		0.783	Sexual Violence
			$\begin{bmatrix} 98 & 6 \\ 4 & 8 \end{bmatrix}$
		0.640	Misogyny-Non-Sexual Violence
			$\begin{bmatrix} 68 & 17 \\ 16 & 15 \end{bmatrix}$
	Best Unimodal Model (T)	**0.788**	Ideological Inequality
			$\begin{bmatrix} 66 & 9 \\ 13 & 18 \end{bmatrix}$
		0.684	Stereotyping-Dominance
			$\begin{bmatrix} 28 & 20 \\ 15 & 53 \end{bmatrix}$
		0.593	Objectification
			$\begin{bmatrix} 81 & 15 \\ 13 & 7 \end{bmatrix}$
		0.732	Sexual Violence
			$\begin{bmatrix} 94 & 10 \\ 4 & 8 \end{bmatrix}$
		0.632	Misogyny-Non-Sexual Violence
			$\begin{bmatrix} 71 & 14 \\ 18 & 13 \end{bmatrix}$

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this chapter, we synthesize the key findings of our research and explore the implications of our results. By addressing our research questions, we draw meaningful conclusions about the distinguishing features of sexist content on TikTok, the performance of AI annotation tools, and the effectiveness of single and multimodal classifiers. Finally, we outline future directions to enhance the detection and categorization of sexism on the platform.

RQ1. What features distinguish sexist TikTok content from non-sexist content, and how do these features contribute to identifying the source intention and categorization of sexism on the platform?

Through our analysis, we have identified significant features that distinguish sexist TikTok content from non-sexist content. These features play a crucial role in identifying the source intention and categorizing the type of sexism present on the platform.

Linguistically, sexist TikToks frequently utilized sexual, affective, and social terms, whereas non-sexist TikToks predominantly featured achievement-related and relational terms. Reported sexism often involved inclusive language and terms linked to social and economic disadvantages, crimes, and disabilities. In contrast, direct sexism was characterized by personal pronouns, terms related to friends, and a higher presence of derogatory language, including references to prostitution, moral defects, and the seven deadly sins. Emotional analysis further differentiated these categories; non-sexist TikToks generally evoked positive emotions such as amusement, joy, and love, whereas sexist TikToks elicited negative responses like anger, disgust, and sadness. Direct sexist content sometimes induced amusement and neutrality, likely due to perceived humor, while reported sexism triggered negative emotions such as disapproval and fear. Significant differences were also noted within specific categories of sexism. For example, content related to sexual violence showed a higher usage of past tense verbs and a stronger expression of anger, while misogynistic content was associated with greater sadness and terms related to death.

In terms of audio features, direct sexist content exhibited higher average of Root Mean Square levels compared to reported sexist content. Reported sexist content, on the other hand, had higher values of spectral contrast and Zero Crossing Rate, which can be attributed to the different audio dynamics and tonal qualities in these videos.

Visual analysis using the BLIP model revealed that certain scenarios and objects in video captions provide clues for detecting sexism or its intention. Terms like 'dark', 'purple', and 'cheerful' were associated with sexist classifications, while neutral terms such as 'dog' or 'road' were linked to non-sexist content. Additionally, 'dark background' and 'smiley' were indicative of direct sexism, whereas 'a woman' and 'social event' pointed towards reported sexism.

RQ2. How well does GPT-3.5 Turbo perform in annotating sexist content on TikTok compared to human annotators, and what is the level of agreement between GPT-3.5 Turbo and human annotators?

The findings from the evaluation of GPT-3.5 Turbo's performance in annotating sexist content on

TikTok suggest that while the AI demonstrates a basic ability to match human annotators, its level of agreement remains lower than desirable for practical application. The fair agreement scores (Cohen's Kappa) for sexism detection and source intention classification tasks indicate that GPT-3.5 Turbo cannot yet replace manual annotation processes, especially when high precision is required. This is why we decided not to consider GPT-3.5 Turbo for annotating TikTok videos.

The significantly lower Cohen's Kappa values in the task of categorizing sexism into five distinct categories further underscore the limitations of using GPT-3.5 Turbo in isolation for complex annotation tasks. Such tasks often require nuanced understanding of context and subtleties in language that current AI technology might not fully capture without additional inputs or training.

This points to a crucial aspect: despite advancements in AI, human oversight remains essential, particularly for tasks involving nuanced social and cultural judgments. The findings highlight the necessity of combining AI tools with human expertise to enhance accuracy and reliability in annotations, rather than expecting AI to fully automate such sensitive processes. The reliance on human annotators cannot be completely eliminated at this stage, especially for content that requires deep contextual and cultural understanding. Thus, while AI like GPT-3.5 Turbo can support and streamline the annotation process, it is not yet capable of completely replacing the costly and labor-intensive manual annotation needed to ensure the high quality and reliability required in studies of social media content, such as sexism on TikTok.

RQ3. How effective are classifiers based on single modalities (text, audio, and video) in detecting sexism, determining source intention, and categorizing different forms of sexism on TikTok? This research question seeks to evaluate the individual strengths and limitations of classifiers focusing on specific modalities.

In our evaluation of classifiers based on single modalities (text, audio, and video), we observed varying levels of effectiveness across different tasks related to detecting and categorizing sexism on TikTok.

For text-based classifiers, embeddings outperformed TF-IDF in capturing semantic relationships. This indicates that embeddings are more effective for understanding the nuances of language used in sexist content. However, incorporating additional linguistic variables did not significantly enhance the performance of text embeddings, suggesting that these variables may be redundant when using advanced embedding techniques.

In contrast, audio and video classifiers did not perform as well for the task of detecting sexism (Task 1) and categorizing different forms of sexism (Task 3). These modalities struggled to match the performance of text-based classifiers, likely due to the more complex and varied nature of audio and visual data.

However, audio and video classifiers showed comparable performance to text embeddings in determining the source intention of sexism (Task 2). Notably, the Vision Transformer model, which averages embeddings from a few frames of video, emerged as the best-performing modality for this task. This suggests that while audio and video may be less effective for direct sexism detection and detailed categorization, they are valuable for understanding the context and intention behind sexist content.

In summary, text-based classifiers, particularly those using embeddings, are highly effective for detecting and categorizing sexism. Audio and video classifiers, while less effective in these areas, provide significant insights into the source intention of sexist content. The integration of these modalities can potentially lead to a more comprehensive and accurate system for identifying and understanding sexism on platforms like TikTok.

RQ4. Do classifiers utilizing a multimodal approach, combining text, audio, and video analysis, outperform single modality classifiers in terms of detecting sexism, understanding source intention, and categorizing different manifestations of sexism on TikTok?

Classifiers that combine text, audio, and video analysis show varying performance across different tasks related to detecting and understanding sexism on TikTok. Initially, using only text outperformed multimodal approaches for detecting the presence of sexism (Task 1). However, a fine-tuned multimodal approach, where the last attention layers of text, audio, and video models were trained together, significantly improved performance, achieving higher F1-macro scores than both the best unimodal and SVM-based multimodal models.

For discerning the intent behind sexist content (Task 2), the multimodal approach demonstrated clear superiority. The TAV model (Text, Audio, and Video) achieved top performance with notable improvements over unimodal models, further enhanced by fine-tuning, which reached F1-macro scores of 0.787 for English and 0.804 for Spanish.

In categorizing different types of sexism (Task 3), the fine-tuning multimodal approach performed well in Spanish, achieving competitive F1-macro scores around 0.700 for all categories, surpassing both unimodal text-based and SVM-based approaches.

In summary, while text-based classifiers excel in detecting sexism, advanced multimodal techniques, particularly fine-tuning across modalities, provide significant benefits in understanding intent and categorizing sexism, highlighting the value of integrating multiple data sources for more accurate analysis.

6.2 Future Work

Future studies should focus on several key areas to enhance the performance and robustness of multimodal models in detecting and understanding sexism on social media platforms like TikTok.

First, expanding and diversifying the datasets is crucial. Increasing the representation of various categories of sexism will enable more accurate and nuanced model training. A diverse dataset, encompassing a wide range of global contexts and demographics, will also improve the generalizability of the models, ensuring they perform well across different cultural and social backgrounds. This approach can help in addressing biases and enhancing the inclusivity of the models.

Second, integrating the innovative paradigm of Learning With Disagreement (LeWiDi) [26] could be highly beneficial. LeWiDi addresses the challenges posed by conflicting annotations, which are common in subjective tasks like sexism detection. By effectively handling these disagreements, models can learn more robustly from diverse perspectives, improving their ability to accurately interpret and classify sexist content. Future work could explore the use of more sophisticated models such as GPT-4[1] or multimodal systems like Gemini 1.5[2], which consider the entire video content. Such advancements could substantially enhance the accuracy and efficiency of automated annotations, potentially reducing the labor-intensive process of manual annotations significantly.

Improving text preprocessing methods is another critical area. Enhanced techniques to avoid information loss during transcription or Optical Character Recognition processes will ensure that the models have access to the most accurate and comprehensive textual data. This step is vital for maintaining the integrity of linguistic features, which are crucial for detecting sexism.

Exploring additional types of features, such as sociolinguistic cues, contextual information, and user interaction patterns, could provide deeper insights into sexist content detection. Incorporating these features could help models better understand the nuances and subtleties of sexist behavior and language. Analyzing the interactions between users and content, including the examination of comments, can provide valuable context and reveal patterns in the spread and reception of sexist content.

Furthermore, investigating the diffusion of sexist content across the TikTok platform can offer critical insights into how such content propagates and gains traction. Understanding the dynamics of content spread, including the role of influential users and network effects, can inform strategies to mitigate the impact of sexist content.

Research should also explore new model architectures, particularly those involving advanced neural networks and attention mechanisms. Some promising approaches include Cross-Modal Attention Networks; these networks allow the model to attend to relevant parts of different modalities simultaneously, improving the integration of textual, auditory, and visual data.

Finally, fostering interdisciplinary collaborations is essential. Bringing together experts from fields such as gender studies, linguistics, sociology, and artificial intelligence can lead to the development of more sophisticated and contextually aware models. These collaborations can drive innovation and ensure that the models are grounded in a deep understanding of gender dynamics and biases.

By focusing on these areas, future research can make substantial contributions towards mitigating gender-based biases and fostering a more inclusive and respectful online environment.

[1] `https://openai.com/index/gpt-4/`

[2] `https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/`

Bibliography

[1] Tosin Adewumi, Sana Sabah Sabry, Nosheen Abid, Foteini Liwicki, and Marcus Liwicki. "T5 for Hate Speech, Augmented Data, and Ensemble". In: *Sci* 5.4 (2023). ISSN: 2413-4155. DOI: 10.3390/sci5040037. URL: https://www.mdpi.com/2413-4155/5/4/37.

[2] Cleber Alcântara, Viviane Moreira, and Diego Feijo. "Offensive video detection: dataset and baseline results". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 4309–4319.

[3] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. "Automatic Identification and Classification of Misogynistic Language on Twitter". In: *Natural Language Processing and Information Systems*. May 2018, pp. 57–64. ISBN: 978-3-319-91946-1. DOI: 10.1007/978-3-319-91947-8_6.

[4] Aymé Arango, Jesus Perez-Martin, and Arniel Labrada. "HateU at SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification". In: *Proceedings of the 16th International Workshop on Semantic Evaluation*. 2022, pp. 581–584. DOI: 10.18653/v1/2022.semeval-1.80.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 12449–12460.

[6] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. "SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter". In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Ed. by Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: https://aclanthology.org/S19-2007.

[7] Elisa Bassignana, Valerio Basile, and Viviana Patti. "Hurtlex: A Multilingual Lexicon of Words to Hurt". In: *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It 2018)*. 2018.

[8] Tom Bourgeade, Patricia Chiril, F. Benamara, and Véronique Moriceau. "What Did You Learn To Hate? A Topic-Oriented Analysis of Generalization in Hate Speech Detection". In: *Proceedings of the 17th International Workshop on Semantic Evaluation*. 2023, pp. 3477–3490. DOI: 10.18653/v1/2023.eacl-main.254.

[9] Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. "ProCap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection". In: *MM '23: The 31st ACM International Conference on Multimedia*. Oct. 2023, pp. 5244–5252. DOI: 10.1145/3581783.3612498.

[10] Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. "Emotionally Informed Hate Speech Detection: A Multi-target Perspective". In: *Cognitive Computation* (2021). DOI: 10.1007/s12559-021-09862-5.

[11] Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. "HateMM: A Multi-Modal Dataset for Hate Video Classification". In: *Proceedings of the International AAAI Conference on Web and Social Media* 17 (June 2023), pp. 1014–1023.

[12] Abreham Gebremedin Debele and Michael Melese Woldeyohannis. "Multimodal Amharic Hate Speech Detection Using Deep Learning". In: *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*. 2022, pp. 102–107. DOI: 10.1109/ICT4DA56482.2022.9971436.

[13] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. "SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification". In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics. 2022, pp. 533–549.

[14] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. "AMI @ EVALITA2020: Automatic Misogyny Identification". In: *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*. Jan. 2020, pp. 21–28. ISBN: 9791280136329. DOI: 10.4000/books.aaccademia.6764.

[15] Francesca Gasparini, Ilaria Erba, Elisabetta Fersini, and Silvia Corchs. "Multimodal Classification of Sexist Advertisements". In: *ICETE 2018 - Proceedings of the 15th International Joint Conference on e-Business and Telecommunications - Volume 2*. SciTePress, Jan. 2018, pp. 399–406. DOI: 10.5220/0006859403990406.

[16] Peter Glick and Susan T. Fiske. "Ambivalent sexism". In: vol. 33. Advances in Experimental Social Psychology. Academic Press, 2001, pp. 115–188. DOI: https://doi.org/10.1016/S0065-2601(01)80005-8. URL: https://www.sciencedirect.com/science/article/pii/S0065260101800058.

[17] Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. "SemEval-2023 Task 10: Explainable Detection of Online Sexism". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2193–2210.

[18] *Linguistic Inquiry and Word Count (LIWC)*. URL: https://www.liwc.app/dictionaries.

[19] Marta Marchiori Manerba and Sara Tonelli. "Fine-Grained Fairness Analysis of Abusive Language Detection Systems with CheckList". In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. 2021. DOI: 10.18653/v1/2021.woah-1.9.

[20] Myron Darrel L. Montefalcon, Jay Rhald C. Padilla, Joshua Lois C. Paulino, Jeline G. Go, Ramon L. Rodriguez, and Joseph Marvin R. Imperial. "Understanding Facial Expression Expressing Hate from Online Short-form Videos". In: *2021 5th International Conference on E-Society, E-Education and E-Technology (ICSET 2021)*. Taipei, Taiwan: ACM, New York, NY, USA, Aug. 2021, pp. 201–207. DOI: 10.1145/3485768.3485785.

[21] Nadim Nachar. "The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution". In: *Tutorials in Quantitative Methods for Psychology* 4 (Mar. 2008). DOI: 10.20982/tqmp.04.1.p013.

[22] John T. Nockleby. *Hate Speech*. Ed. by et al. Leonard W. Levy Kenneth L. Karst. 2nd ed. New York: Macmillan, 2000, pp. 1277–1279.

[23] Endang Wahyu Pamungkas, Valerio Basile, and V. Patti. "Misogyny Detection in Twitter: a Multilingual and Cross-Domain Study". In: *Information Processing & Management* 57 (2020), p. 102360. DOI: 10.1016/j.ipm.2020.102360.

[24] Jayant Panwar and Radhika Mamidi. "PanwarJayant at SemEval-2023 Task 10: Exploring the Effectiveness of Conventional Machine Learning Techniques for Online Sexism Detection". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1531–1536. DOI: 10.18653/v1/2023.semeval-1.211. URL: https://aclanthology.org/2023.semeval-1.211.

[25] Andrei Paraschiv, Mihai Dascalu, and Dumitru-Clementin Cercel. "UPB at SemEval-2022 Task 5: Enhancing UNITER with Image Sentiment and Graph Convolutional Networks for Multimedia Automatic Misogyny Identification". In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Ed. by Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 618–625. DOI: 10.18653/v1/2022.semeval-1.85. URL: https://aclanthology.org/2022.semeval-1.85.

[26] Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. "Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, and Nicola Ferro. Cham: Springer Nature Switzerland, Sept. 2023, pp. 316–342. ISBN: 978-3-031-42448-9. DOI: 10.1007/978-3-031-42448-9_23.

[27] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. "Robust Speech Recognition via Large-Scale Weak Supervision". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, 23–29 Jul 2023, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

[28] Aneri Rana and Sonali Jha. "Emotion Based Hate Speech Detection using Multimodal Learning". In: *ArXiv* abs/2202.06218 (2022). URL: https://api.semanticscholar.org/CorpusID:246822635.

[29] Rutuja G. Rathod, Y. Barve, Jatinderkumar R. Saini, and Sourav Rathod. "From Data Preprocessing to Hate Speech Detection: An Interdisciplinary Study on Women-targeted Online Abuse". In: *2023 3rd International Conference on Intelligent Technologies (CONIT)*. 2023, pp. 1–8. DOI: 10.1109/CONIT59222.2023.10205571.

[30] Rakib Hossain Rifat, A. Shruti, Marufa Kamal, and Farig Sadeque. "ACSMKRHR at SemEval-2023 Task 10: Explainable Online Sexism Detection(EDOS)". In: *Proceedings of the 17th International Workshop on Semantic Evaluation*. 2023, pp. 724–732. DOI: 10.18653/v1/2023.semeval-1.99.

[31] F. Rodríguez-Sánchez et al. "Overview of EXIST 2021: Sexism identification in social networks". In: *Procesamiento del Lenguaje Natural* 67 (2021), pp. 195–207.

[32] F. Rodríguez-Sánchez et al. "Overview of EXIST 2022: Sexism identification in social networks". In: *Procesamiento del Lenguaje Natural* 69 (2022), pp. 229–240.

[33] F. Rodríguez-Sánchez, Jorge Carrillo-de-Albornoz, and Laura Plaza. "Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data". In: *IEEE Access* 8 (2020), pp. 219563–219576. DOI: 10.1109/ACCESS.2020.3042604.

[34] Rajalakshmi Sivanaiah, S. AngelDeborah, S. M. Rajendram, and T. T. Mirnalinee. "TechSSN at SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification using Deep Learning Models". In: *Proceedings of the 16th International Workshop on Semantic Evaluation*. 2022, pp. 571–574. DOI: 10.18653/v1/2022.semeval-1.78.

[35] Fariha Tahosin, Ponkoj Shill, and Md Golam Rabiul Alam. "Multi-modal Hate Speech Detection using Machine Learning". In: Dec. 2021, pp. 4496–4499. DOI: 10.1109/BigData52589.2021.9671955.

[36] Surendrabikram Thapa, Farhan Ahmad Jafri, Ali Hürriyetoğlu, Francielle Vargas, Roy Ka Wei Lee, and Usman Naseem. "Multimodal Hate Speech Event Detection - Shared Task 4, Case 2023". In: *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*. Incoma Ltd, 2023, pp. 151–159.

[37] Ching Seh Wu and Unnathi Bhandary. "Detection of hate speech in videos using machine learning". In: *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE. 2020, pp. 585–590.

[38] Mutaz Younes, Ali Kharabsheh, and Mohammad Bani Younes. "Alexa at SemEval-2023 Task 10: Ensemble Modeling of DeBERTa and BERT Variations for Identifying Sexist Text". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1644–1649. DOI: 10.18653/v1/2023.semeval-1.228. URL: https://aclanthology.org/2023.semeval-1.228.

Appendix A

Results on Sexism in Spanish

Table A.1: LIWC Spanish 2007 Features

Table A.2: Linguistic Processes

Category	Feature Example	Length
Funct	durante	443
TotPron	vos	68
PronPer	yo	33
Yo	me	6
Nosotro	nos	3
TuUtd	contigo	11
ElElla	le	11
Ellos	suyo*	9
PronImp	algún	35
Articulo	un	10
Verbos	sufrirá*	7179
VerbAux	podríais	99
Pasado	brillába*	3341
Present	trato	2812
Futuro	huiré*	1022
Adverb	relativamente	160
Prepos	de	36
Conjunc	ni	30
Negacio	negación	16
Cuantif	secciones	83
Numeros	seiscient*	69
Maldec	mamaron	110
verbYO	puedo	104
verbTU	estarías	118
verbNOS	pedíamos	101
verbosEL	haya	106
verbELLOS	contaron	104
Subjuntiv	comunicásemos	2020
VosUtds	vosotr*	5
formal	ustedes	2
informal	tuya*	8
verbVos	creeréis	103

Table A.3: Psychological Processes

Category	Feature Example	Length
Social	donase*	1532
Familia	matern*	50
Amigos	amiga	45
Humanos	hembra*	55
Afect	perdí	1822
EmoPos	aliento	695
EmoNeg	perdí	1087
Ansiedad	sobresaltó	135
Enfado	hostilidad*	506
Triste	extrañó	262
MecCog	preguntamos	3212
Insight	preferible*	1485
Causa	hacerse	487
Discrep	prefirieron*	150
Tentat	supusimos	383
Certeza	garantía*	128
Inhib	evitéis	333
Incl	también	28
Excl	ni	35
Percept	ásper*	1387
Ver	vigilar	340
Oir	callar	433
Sentir	caricia*	326
Biolog	botana*	1026
Cuerpo	peso	370
Salud	linfoma	219
Sexual	desnud*	191
Ingerir	probó	353

Table A.4: Personal Concerns

Category	Feature Example	Length
Relativ	cerraremos	1943
Movim	caminando	1246
Espacio	poco*	359
Tiempo	retirad*	458
Trabajo	meser*	928
Logro	espléndid*	929
Placer	actuáis	417
Hogar	televis*	146
Dinero	presto	299
Relig	salmo	288
Muerte	cuerpo	130

Table A.5: Spoken Categories

Category	Feature Example	Length
Asentir	afirmaciones	47
NoFluen	ah	8
Relleno	unpoco	6

Table A.6: Significant Differences in EmoRoberta Emotions for Sexist vs Non-Sexist Spanish TikToks

Emotion	**p-value**	**Non-sexist Mean**	**Non-sexist SD**	**Sexist Mean**	**Sexist SD**
admiration	0.0012	0.0478	0.1813	**0.0483**	0.1880
amusement	0.0108	**0.0429**	0.1752	0.0359	0.1639
anger	2.75×10^{-6}	0.0279	0.1274	**0.0465**	0.1674
annoyance	1.32×10^{-11}	0.0178	0.0868	**0.0295**	0.1118
desire	0.0057	**0.0102**	0.0765	0.0091	0.0748
disappointment	2.76×10^{-9}	0.0093	0.0580	**0.0122**	0.0647
disapproval	1.67×10^{-14}	0.0177	0.0896	**0.0326**	0.1285
disgust	1.22×10^{-9}	0.0092	0.0715	**0.0151**	0.0923
embarrassment	9.83×10^{-6}	0.0034	0.0316	**0.0067**	0.0672
excitement	9.82×10^{-6}	**0.0102**	0.0676	0.0082	0.0693
gratitude	0.0087	0.0247	0.1421	**0.0266**	0.1434
joy	1.34×10^{-7}	**0.0349**	0.1531	0.0211	0.1179
love	0.0078	**0.0661**	0.2191	0.0330	0.1517
realization	7.39×10^{-6}	0.0327	0.1406	**0.0484**	0.1632
sadness	0.0008	0.0243	0.1150	**0.0257**	0.1161

Table A.7: Significant Differences in EmoRoberta Emotions between Reported and Direct Sexism Spanish TikToks

Emotion	**p-value**	**Reported Mean**	**Reported SD**	**Direct Mean**	**Direct SD**
amusement	0.0022	0.0211	0.1328	**0.0430**	0.1766
anger	0.0081	**0.0480**	0.1684	0.0458	0.1670
caring	0.0029	**0.0556**	0.1867	0.0288	0.1214
disappointment	5.33×10^{-6}	**0.0179**	0.0810	0.0094	0.0550
disapproval	0.0002	**0.0478**	0.1578	0.0253	0.1112
disgust	0.0096	**0.0210**	0.1032	0.0122	0.0865
embarrassment	0.0021	**0.0078**	0.0654	0.0062	0.0682
fear	0.0004	**0.0186**	0.1055	0.0102	0.0708
grief	2.75×10^{-7}	**0.0062**	0.0513	0.0022	0.0194
nervousness	0.0040	0.0016	0.0096	**0.0032**	0.0297
optimism	0.0188	**0.0074**	0.0405	0.0062	0.0454
realization	4.55×10^{-5}	**0.0638**	0.1896	0.0409	0.1484
relief	0.0004	0.0008	0.0043	**0.0016**	0.0123
remorse	1.42×10^{-7}	**0.0138**	0.0796	0.0072	0.0618
sadness	3.21×10^{-5}	**0.0429**	0.1518	0.0175	0.0932
neutral	0.0040	0.3335	0.3812	**0.3881**	0.3882

Appendix B

Sexism Identification on TikTok: A Multimodal AI Approach with Text, Audio, and Video

Iván Arcos and Paolo Rosso. Sexism Identification on TikTok: A Multimodal AI Approach with Text, Audio, and Video.
In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, LNCS, Springer.

This appendix documents the acceptance of the paper titled "Sexism Identification on TikTok: A Multimodal AI Approach with Text, Audio, and Video" by Iván Arcos and Paolo Rosso:

`https://clef2024.imag.fr/index.php?page=Pages/accepted_papers.html`

for presentation at the CLEF 2024 conference in Grenoble, France, from 9-12 September 2024.

Abstract. Sexism persists as a pervasive issue in society, particularly evident on social media platforms like TikTok. This phenomenon encompasses a spectrum of expressions, ranging from subtle biases to explicit misogyny, posing unique challenges for detection and analysis. While previous research has predominantly focused on textual analysis, the dynamic nature of TikTok demands a more comprehensive approach. This study leverages advancements in Artificial Intelligence (AI), specifically multimodal deep learning, to establish a robust framework for identifying and interpreting sexism on TikTok. We compiled the first dataset of TikTok videos tailored for analyzing sexism in both English and Spanish. This dataset serves as an initial benchmark for comparing models or for future investigations in this area. By integrating text, linguistic features, emotions, audio, and video features, this study identifies unique indicators of sexist content. Multimodal analysis surpasses text-only methods, particularly in understanding the intentions behind sexism.

Keywords – Multimodal Sexism Identification, TikTok, Artificial Intelligence.

Structure. The rest of the paper is structured as follows. Section 2 presents some related work. Section 3 introduces the tasks of sexism detection, source intention classification and sexism categorization, as well as the the dataset we compiled. Section 4 describes the models for text, audio, video and multimodal data. Section 5 presents the results and, finally, Section 6 draws some conclusions and discusses future work.

Appendix C

Final Project's Contribution to Sustainable Development Goals

Table C.1: Relationship of the project with the Sustainable Development Goals (SDGs) of the 2030 Agenda

SDG	Description	High	Medium	Low	No Proceeds
1	No Poverty				✓
2	Zero Hunger				✓
3	Good Health and Well-being	✓			
4	Quality Education	✓			
5	Gender Equality	✓			
6	Clean Water and Sanitation				✓
7	Affordable and Clean Energy				✓
8	Decent Work and Economic Growth				✓
9	Industry, Innovation and Infrastructure				✓
10	Reduced Inequalities	✓			
11	Sustainable Cities and Communities				✓
12	Responsible Consumption and Production				✓
13	Climate Action				✓
14	Life Below Water				✓
15	Life on Land				✓
16	Peace, Justice, and Strong Institutions	✓			
17	Partnerships for the Goals				✓

SDG 3: Good Health and Well-being

The project targets SDG 3 by deploying AI tools to detect and reduce sexist content on TikTok, thus decreasing the mental and emotional strain on users. This initiative is pivotal in promoting a more supportive online environment that upholds the well-being of users, especially those who might be more vulnerable to the impacts of online harassment.

SDG 4: Quality Education

This project contributes to SDG 4 by using insights from social media analysis to develop educational materials that address gender biases. These materials can be integrated into school curricula, helping to educate students about digital responsibility and the importance of respecting gender diversity online. This promotes a well-rounded educational experience that extends beyond traditional learning environments.

SDG 5: Gender Equality

By focusing on identifying and mitigating sexist content on TikTok, this project directly supports SDG 5. It not only helps in moderating harmful content but also serves as a platform for educating content creators and users about gender equality. This dual approach helps foster a digital culture that respects and promotes gender balance.

SDG 10: Reduced Inequality

The project enhances SDG 10 efforts by making TikTok a more inclusive platform, where discriminatory content is actively identified and reduced. By leveling the digital playing field, the project ensures that all individuals, regardless of gender, have equitable social media experiences. This helps to prevent the perpetuation of existing social inequalities and promotes inclusivity at a broader scale.

SDG 16: Peace, Justice, and Strong Institutions

In alignment with SDG 16, the project improves peace and justice on TikTok by developing and implementing AI-driven moderation tools that effectively identify and curb online violence and harassment. These technologies strengthen the platform's regulatory framework, contributing to safer and more respectful interactions, which are crucial for maintaining social order and justice in digital communities.

Table C.2: Relación del proyecto con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030

ODS	**Descripción**	**Alto**	**Medio**	**Bajo**	**No Procede**
1	Fin de la pobreza				✓
2	Hambre cero				✓
3	Salud y bienestar	✓			
4	Educación de calidad	✓			
5	Igualdad de género	✓			
6	Agua limpia y saneamiento				✓
7	Energía asequible y no contaminante				✓
8	Trabajo decente y crecimiento económico				✓
9	Industria, innovación e infraestructura				✓
10	Reducción de las desigualdades	✓			
11	Ciudades y comunidades sostenibles				✓
12	Producción y consumo responsables				✓
13	Acción por el clima				✓
14	Vida submarina				✓
15	Vida de ecosistemas terrestres				✓
16	Paz, justicia e instituciones sólidas	✓			
17	Alianzas para lograr los objetivos				✓

ODS 3: Salud y bienestar

El proyecto se enfoca en el ODS 3 utilizando herramientas de inteligencia artificial para detectar y reducir el contenido sexista en TikTok, disminuyendo así el estrés mental y emocional de los usuarios. Esta iniciativa es fundamental para fomentar un ambiente en línea más acogedor que proteja el bienestar de los usuarios, especialmente de aquellos más vulnerables a los efectos del acoso en línea.

ODS 4: Educación de calidad

Este proyecto aporta al ODS 4 al aprovechar análisis de redes sociales para crear materiales educativos que enfrenten los sesgos de género. Estos recursos se pueden incorporar en los planes de estudio escolares, capacitando a los estudiantes en responsabilidad digital y en la importancia de valorar la diversidad de género en internet. Esto enriquece la experiencia educativa, llevándola más allá de los entornos tradicionales de aprendizaje.

ODS 5: Igualdad de género

Centrándose en la identificación y mitigación del contenido sexista en TikTok, este proyecto respalda directamente el ODS 5. No solo contribuye a moderar contenido perjudicial, sino que también educa a creadores y usuarios sobre la igualdad de género. Este enfoque combinado promueve una cultura digital que valora y fomenta la igualdad de género.

ODS 10: Reducción de las desigualdades

El proyecto contribuye al ODS 10 haciendo de TikTok un espacio más inclusivo, donde se identifica y se reduce el contenido discriminatorio activamente. Al equilibrar las condiciones en el entorno digital, asegura que todas las personas, sin importar su género, disfruten de experiencias equitativas en las redes sociales. Esto evita la perpetuación de desigualdades sociales y fomenta una inclusividad más amplia.

ODS 16: Paz, justicia e instituciones sólidas

De acuerdo con el ODS 16, el proyecto fortalece la paz y la justicia en TikTok mediante el desarrollo e implementación de herramientas de moderación basadas en IA, que identifican y controlan efectivamente la violencia y el acoso en línea. Estas tecnologías refuerzan el marco normativo de la plataforma, asegurando interacciones más seguras y respetuosas, esenciales para mantener el orden social y la justicia en las comunidades digitales.