MISSIONE 4
ISTRUZIONE
RICERCA

Department of **Informatics, Systems and Communication**

Ph. D. program in **Computer Science**, **XXXVIII** cycle

Supervised and Semi-Supervised Methods for Energy-Efficient Multimodal Hate Detection

Fariha Maqbool
Registration number: 906926

Supervisor: **Prof. Elisabetta Fersini**
Tutor: **Prof. Gianluigi Ciocca**
PhD Program Director: **Prof. Gianluca Della Vedova**

Academic Year **2025/2026**

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Abstract

The rapid growth of user-generated content on social media has increased the spread of hateful content, posing significant challenges for content moderation systems. Among these forms of abuse, sexism and misogyny have emerged as a particularly pervasive and damaging phenomenon. Detecting misogynistic and sexist content is further complicated by the multimodal nature of online communication, especially memes, as well as by the inherent subjectivity and disagreement among human annotators when labeling such content. Although recent advances in large multimodal models have demonstrated strong performance in this area, but they require substantial computational resources and have a significant environmental impact, limiting their practicality for large-scale deployment. Moreover, prevailing approaches treat annotation disagreement as noise to be suppressed rather than as meaningful signal about content ambiguity. This thesis investigates computationally efficient and environmentally sustainable approaches for multimodal sexism and misogyny detection, that explicitly model annotation disagreement. Grounded in the principles of Green AI, the work explores lightweight methods that leverage pretrained representations without relying on extensive fine-tuning or large labeled datasets. A semi-supervised constrained clustering method is first introduced to leverage pretrained vision–language embeddings with minimal annotation and computational cost, achieving competitive performance while offering favorable energy–accuracy trade-offs. However, this method assumes unambiguous ground truth labels, overlooking the contested nature of some content. The second stage develops a supervised contrastive learning framework that detects hate while simultaneously modeling annotator disagreement as a complementary task. Experiments show that the same lightweight architecture performs effectively on both hate and disagreement detection, though the latter proves to be inherently more challenging, reflecting the difficulty of predicting human perceptual variation. This framework is also extended to joint hate-disagreement prediction, enabling a single efficient model to simultaneously detect hateful content and flag ambiguous cases requiring human review. Together, these findings demonstrate that robust, uncertainty-aware content moderation systems can be built without reliance on large, resource-intensive models, and that modeling subjectivity is not only feasible but also compatible with computational efficiency.

List of publications

[1] **Fariha Maqbool** and Elisabetta Fersini. A contrastive learning based approach to detect sexism in memes. In Guglielmo Faggioli, Nicola Ferro, Petra Galuscáková, and Alba García Seco de Herrera, editors, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of CEUR Workshop Proceedings, pages 1091–1097

[2] **Fariha Maqbool** and Elisabetta Fersini. Multimodal hate speech detection in memes from mexico using BLIP. In Salud María Jiménez-Zafra, Luis Chiruzzo, Francisco Rangel, Fazlourrahman Balouchzahi, Ulisses Brisolara Corrêa, Alba Bonet-Jover, Helena Gómez-Adorno, José Ángel González Barba, Delia Irazú Hernández Farías, Arturo Montejo-Ráez, Pablo Moral, Carlos Rodríguez Abellán, María Estrella Vallecillo Rodríguez, Mariona Taulé, and Rafael Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Valladolid, Spain, September 24, 2024*, volume 3756 of CEUR Workshop Proceedings.

[3] Paolo Italiani, **Fariha Maqbool**, David Gimeno-Gómez, Elisabetta Fersini, and Carlos-D Martínez-Hinarejos. Trankiltwice at exist2025: detecting sexism in memes under multi-lingual settings. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, 2025.

Submitted Research Papers:

[1] **Fariha Maqbool**, Paolo Rosso, and Elisabetta Fersini. A Semi-Supervised Clustering Approach for Energy-Efficient Multimodal Misogyny and Sexism Detection. Submitted to *Online Social Networks and Media (OSNEM)*.

[2] **Fariha Maqbool**, Giulia Rizzi, and Elisabetta Fersini. Modeling Misogyny and Perspectivism in Memes: A Contrastive Learning Approach. Submitted to *Multimedia Tools and Applications*.

[3] Giulia Rizzi, **Fariha Maqbool**, and Elisabetta Fersini. Hate against Women: A survey Submitted to *Information Processing and Management*.

Acknowledgements

I would first like to express my deepest gratitude to my supervisor, Prof. Elisabetta Fersini, whose constant support has guided me through every stage of this journey. She understood the challenges I faced as an international student and supported me at every level. Her mentorship and encouragement allowed me to grow academically and professionally, and I am truly grateful for the many opportunities she gave me to learn and develop. I am also grateful to Prof. Paolo Rosso for welcoming me during my research stay in Valencia and for the guidance that made that period such a meaningful learning experience.

I want to thank my colleagues at Bicocca and Valencia for everything I learned from them and for their support along the way.

My heartfelt thanks go to my parents and siblings, whose unwavering support made it possible for me to leave my home country and pursue this doctoral degree. Their emotional strength and belief in me have been a constant source of motivation.

Finally, I would like to thank my friends especially those at the U62 Bicocca residence, who have become like family to me. They stood by me through both the good moments and the difficult ones, and this journey would not have been the same without their companionship and support. A special thanks to Rabia Maryam, who was always there to help and support me in every situation.

I acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing, under the ISCS project funded by the NextGenerationEU – CUP H43C22000520001.

Contents

List of Figures

List of Tables

1 Introduction

Social media have become an essential part of online communication in the modern era, allowing users to freely publish and share their thoughts, emotions and opinions at any time. However, the exponential growth in user-generated content has also amplified the spread of hate speech. Among the most critical issues emerging in this context is **sexism** [51], a type of gender-based bias that promotes the belief that one gender is superior to another. Within this broader category, **misogyny** represents a significant form of hate specifically targeting women [118]. Misogyny can be seen as a stronger and more extreme form of sexism. It reflects a deeper sexist attitude that becomes part of people's thinking. It can also be understood as an organized or intensified form of sexism, especially because it involves feelings of hostility and hate [78]. It has been shown to harm women's self-esteem, cognitive performance, and career ambitions, while reinforcing traditional gender roles and dependency behaviors [75].

The scale and rapid pace of online content generation makes it impractical to rely solely on human moderation; therefore, it requires the development of automated methods capable of detecting and detoxifying inappropriate language. Consequently, a growing body of research has focused on understanding and identifying misogyny and sexism in online spaces. Much of this work has concentrated on textual abuse, which remains one of the most prevalent forms of online sexism and misogyny and has gained increasing relevance in recent years [7, 33]. However, the challenge of sexism and misogyny detection extends beyond purely textual data. In multimodal content such as memes, the complexity of the problem increases considerably [48]. Yet these automated systems face a fundamental challenge: Hate and offensive content are not always clearly defined or universally agreed upon. What is considered hateful or harmful often depends on the perspective of the person interpreting it. Individual background, cultural norms, personal experiences, and social context all shape how people perceive language and images. As a result, the same content may be seen as offensive by some, neutral or humorous by others, and even acceptable within certain communities. This inherent subjectivity makes the identification of hate particularly challenging. **Disagreement** is a natural outcome when people with diverse viewpoints evaluate the same content, and such differences do not necessarily indicate error but rather reflect the complexity of human judgment. Recognizing this subjectivity is essential when studying or addressing hate, as it

highlights the importance of considering multiple perspectives rather than assuming a single, fixed interpretation.

The subjectivity of hateful content has important consequences for how automated detection systems are designed. The prevailing response in the field has been to scale up model size and training data, treating annotation disagreement as label noise to be overcome through sheer capacity. However, this approach is both scientifically limiting and computationally costly because it discards meaningful information about content ambiguity and produces systems that are impractical for large-scale deployment. This thesis takes a different position: disagreement among annotators is not noise but a signal, reflecting genuine ambiguity in the content itself. Building efficient systems that acknowledge this subjectivity rather than suppress it is both a scientific and a practical necessity.

Recent advancements using neural architectures have shown greater promise in handling the complexities of hate and disagreement. Recently, Large Vision-Language Models (LVLMs) have been proven effective but they rely on billions of parameters and require substantial computational resources for both training and inference [131]. This results in high energy consumption and significant carbon emissions, making such approaches impractical and costly for large-scale or real-time content moderation systems.

The primary goal of this thesis is to develop computationally efficient and environmentally sustainable methods for multimodal hate detection that explicitly account for the subjective nature of hateful content. Grounded in the principles of **Green AI**, this work demonstrates that robust content moderation does not require large resource-intensive models, and that annotator disagreement can be modeled within lightweight frameworks that operate on pre-trained embeddings. This is achieved through a two-stage investigation: first, establishing that classical clustering methods with minimal modifications can achieve competitive hate detection with favorable energy-performance trade-offs. Second, demonstrating that the same efficiency principles extend to the harder problem of predicting content ambiguity: whether a given instance is likely to evoke disagreement among human annotators.

1.1 Thesis Contributions

The work carried out in this thesis focuses on multimodal sexism, misogyny and disagreement detection by (1) proposing lightweight and computationally efficient frameworks, and (2) systematically analyzing their effectiveness and sustainability. The main contributions of the thesis are summarized as below:

- A constrained clustering framework based on classical clustering methods is proposed, introducing a regularization strategy in the embedding space to effectively exploit the

representational power of pretrained models. The study further examines how progressively increasing the amount of labeled data affects both performance and energy usage, revealing the trade-offs that emerge when moving from limited to fully supervised training. This framework establishes that competitive hate detection is achievable with minimal computational cost, but treats annotation labels as unambiguous ground truth, a limitation that motivates the explicit modeling of disagreement in the subsequent contribution.

- Building on the efficiency principles established through clustering, a supervised contrastive learning framework is introduced that addresses both hate detection and disagreement detection. The framework treats disagreement detection as a classification task: predicting whether annotators will agree or disagree on a given piece of content. Extensive evaluation across pretrained encoders, fusion strategies, and aggregation methods reveals that the same lightweight architecture performs best on both tasks, though disagreement detection proves fundamentally harder, reflecting the challenge of predicting human perceptual ambiguity. The framework is further extended to joint prediction, enabling a single model to simultaneously detect misogynistic content and identify ambiguous cases requiring human review.
- A detailed analysis of the trade-off between model performance and carbon energy consumption is conducted, demonstrating the efficiency of the proposed methods compared to baseline approaches. Throughout both contributions, computational efficiency is treated not merely as a desirable property but as a design constraint.

1.2 Organization of Thesis

The thesis is organized as follows:

- **CHAPTER 1: INTRODUCTION.** This chapter introduces the overall scope of the thesis, presenting the central themes and key concepts that guide the research. It also outlines the main contributions and provides a roadmap for the remainder of the work.
- **CHAPTER 2: PRELIMINARIES.** This chapter presents the foundational concepts relevant to the study, with particular focus to misogyny, sexism, annotator disagreement, and Green AI. It further offers an overview of multimodal AI and contrastive learning, which are the core part of the thesis.
- **CHAPTER 3: LITERATURE REVIEW.** This chapter surveys the state of the art in hate speech detection, with a specific focus on misogyny and sexism. It also reviews

existing approaches within the Learning with Disagreements paradigm and examines current Green AI methodologies.

- **CHAPTER 4: DATASETS AND EVALUATION METHODS.** This chapter describes the datasets employed throughout the experimental work and details the evaluation methodologies used to assess model performance. It also provides an overview of the vision–language models considered in the study.
- **CHAPTER 5: SEMI-SUPERVISED CLUSTERING FOR SEXISM AND MISOGYNY DETECTION.** This chapter explores a constrained clustering strategy designed to address the challenges of misogyny and sexism detection. The approach leverages the representational strength of large pretrained models while minimizing the need for extensive labeled data or computationally expensive training.
- **CHAPTER 6: CONTRASTIVE LEARNING BASED METHOD FOR IDENTIFYING HATE AND DISAGREEMENT.** This chapter presents the proposed framework and the methodological innovations developed to tackle multimodal misogyny detection and annotator disagreement. It further extends the method to a joint Hate–Disagreement prediction task and discusses the associated challenges.
- **CHAPTER 7: CONCLUSIONS AND FUTURE WORKS.** This thesis concludes with a summary of the main findings and an overview of the most significant contributions to research in this field. It also suggests directions for future research that could build upon the findings presented in this thesis.

2 Preliminaries

In this chapter, we introduce the foundations of this PhD thesis in order to give the reader most of the required knowledge to engage the rest of this thesis. This chapter starts with an overview of the three main concepts addressed within this thesis: hate, disagreement, and green AI. Then, we also provide a broader view of multimodal AI, which constitutes the core of the data used in this thesis. We also provide an overview of the contrastive learning, which is one of the key methodological approaches employed in our research.

2.1 Hate Speech

The term 'Hate Speech' was introduced by a group of legal scholars in the United States during the late 1980s, who sought to describe and compare how different legal systems addressed forms of harmful racist expression [23]. Hate speech is widely recognized as a complex phenomenon, intrinsically associated with relationships between groups, and also relying on language nuances. Although it has long been a persistent social issue, its forms and modes of expression have evolved significantly over time, particularly with the rise of digital communication platforms. As a result, defining hate speech has become a central challenge in both academic research and policy development.

Despite extensive scholarly attention, there is still no universally accepted definition of hate speech. Many authors highlight that existing definitions are often vague, inconsistent, or contradictory [59]. For instance, [9] emphasize the absence of a single agreed-upon definition. Table 2.1 summarizes several influential definitions proposed by institutions and researchers. In this thesis, we adopt the common conceptual thread shared across these definitions, understanding hate speech as expressions that target, demean, or harm individuals or groups on the basis of assigned or perceived group characteristics such as ethnicity, nationality, gender, sexual orientation, disability, or religion.

Following the definitions presented above, it is important to consider the broader environment in which hate speech circulates. It has been growing in recent years, not only in face-to-face interactions but also in online communication. This issue is made worse by social media platforms, which facilitate the quick dissemination of offensive, profane and obscene

Source	Definition
[160]	Public shouting at someone contrary to good morals and private defamation.
[104]	Language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, or sexual orientation/gender identity.
[67]	Hate speech is a derogatory expression (e.g., words, posts, text messages, images, videos) about people (directly or vicariously) on the basis of assigned group characteristics (e.g., ethnicity, nationality, gender, sexual orientation, disability, religion), based on an intention to harm and with the potential to cause harm at individual, communal, or societal levels.
[119]	Any kind of expression that targets individuals or groups based on their gender, sexual orientation, race, religion, ethnicity, or nationality.
[145]	Any kind of communication in speech, writing, or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group based on their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factor.

Table 2.1: Definitions of Hate Speech from Literature

language, among other inappropriate content. This content spreads through various means, such as text, images, multimedia, and other forms of digital communication. Despite the negative nature of this content, it unfortunately possesses certain qualities that contribute to its rapid dissemination.

In light of this growing digital exposure, the harms associated with hate speech become even more significant. This type of speech can incite violence, promote prejudice, and cause various other harmful effects on individuals and communities. The consequences faced by victims of targeted hate speech are not limited to physical harm; they also experience a profound sense of dread and rejection within their communities. Recognizing the urgency to create online spaces free of racism and hate speech, researchers emphasize the importance of early detection mechanisms to mitigate the pervasive harm caused by such content[50].

2.1.1 Sexism and Misogyny

Sexism is a type of bias and prejudice that leads to harmful sex-based stereotypes and societal expectations. It often involves a combination of gender-based beliefs, attitudes, and actions that result in uneven treatment of men and women. It often promotes the belief that one gen-

der is superior to another. Although it can affect individuals of any gender, women are most often its primary targets [42]. This bias is deeply rooted in history and culture, shaped by long-standing ideas of male superiority, and continues to influence many areas of life, including the workplace, politics, society, and the family [43]. Such experiences have been shown to harm women's self-esteem, cognitive performance, career ambitions, and encourage traditional gender roles and dependency behaviors [75].

Sexism can take several forms, including blatant, covert, and subtle sexism [16]. Blatant sexism refers to obvious and intentional unequal treatment of women, whereas covert sexism involves similarly intentional discrimination that is deliberately concealed. In contrast, subtle sexism consists of unequal treatment that often goes unrecognized because it appears normative. It is hidden like covert sexism, but it is not intentionally harmful. Subtle sexism is considered especially important because it is widespread [16] and can have insidious effects on its targets [141].

Within this broader category of sexism, **Misogyny** represents a significant form of hate specifically targeting women [118]. The term misogyny has historically referred to "hate or contempt for women," a meaning recorded as early as 1656. Its etymology is direct: the Greek miso means "hatred," and gune means "woman" [153, 65]. Over time, dictionary definitions have evolved. In 2002, the Oxford English Dictionary broadened its definition from "hatred of women" to "a feeling of hate or dislike towards women, or a feeling that women are not as good as men," while the American Heritage Dictionary defines misogyny as "hatred or mistrust of women" [65]. Scholars further describe misogyny as "a cultural disposition of hostility towards women solely based on their gender" [66].

With the rise of social networks, long-standing hostility toward women has found new forms of expression in digital communication. Although online spaces have enhanced women's voices, they have also become fertile ground for the spread of misogynistic content. The online environment enables misogyny to appear in many linguistic forms, from subtle exclusion and discrimination to more explicit expressions such as sexual objectification and violent threats [49]. Such content is constantly spreading online and affecting the safety and well-being of women across various platforms. According to a report, women are more affected by online harassment compared to other groups [44]. A related study in the United States showed that 33% of women under 35 have experienced sexual harassment online, in contrast to 11% of men [148]. Another report noted that although interpersonal violence can affect individuals of any gender, women are much more likely to experience severe types of abuse, both in digital spaces and in everyday life [143]. Overall, these findings highlight the persistent and evolving nature of sexism and misogyny across modern social and digital contexts.

2.1.2 Disagreement

Modern machine learning (ML) systems rely heavily on large datasets annotated by humans, whether through crowdsourcing platforms or spontaneous online interactions. Traditional annotation practices typically assume that each instance has a single correct label called *gold standard*. However, this assumption oversimplifies how humans perceive and categorize complex and subjective phenomena. In practice, annotation projects frequently encounter **Disagreement**, which may arise from accidental mistakes, misunderstandings, task ambiguity, or the annotators' subjective beliefs [105].

This issue is particularly pronounced in hate speech detection, where definitions are inherently subjective. Differences in cultural backgrounds, personal experiences, and contextual interpretations lead annotators to classify the same content in different ways. What one annotator perceives as hateful may be considered non-hateful by another. Such variability highlights the importance of acknowledging multiple viewpoints rather than imposing a single "correct" interpretation.

Recognizing this challenge, researchers have begun to question the classical gold--standard paradigm and explore alternative ways of defining ground truth. **Perspectivism** [52] proposes preserving annotations from multiple individuals to model the diversity of human judgments. Instead of collapsing disagreement into a single label, perspectivist approaches retain soft labels that reflect the distribution of annotator opinions, thereby capturing the range of interpretations present in the data.

Building on this idea, the **Learning With Disagreement (LeWiDi)** paradigm [144] shifts the focus from majority voting to models that explicitly learn from diverse perspectives. Rather than discarding disagreements, LeWiDi incorporates all annotator interpretations into the learning process, enabling models to better reflect the complexity of subjective tasks such as hate speech detection.

This broader recognition of subjectivity has encouraged the development of new ground-truth frameworks. Recent work proposes a spectrum ranging from the traditional gold standard to a "diamond standard," where multiple labels are preserved throughout the entire ML pipeline. Empirical studies further show that training on soft labels can outperform training on aggregated labels, particularly when datasets are large and annotators are reliable. Similar findings in hate speech classification demonstrate that models informed by multiple annotator perspectives achieve better performance than those trained solely on fully aggregated labels [105].

2.2 Green AI

The term **Green AI** refers to AI research that yields novel results while explicitly considering the computational cost, encouraging a reduction in the resources required to achieve those results [133]. This definition provides a foundation for understanding the growing concern around the environmental impact of modern machine learning systems.

Over the past decade, artificial intelligence (AI) and machine learning (ML) have transformed numerous sectors including healthcare, finance, transportation, education, and entertainment, by significantly improving efficiency and accuracy. Achieving these performance gains has required increasingly complex models with rapidly growing numbers of parameters. These state-of-the-art NLP and machine learning models require large quantities of energy and water to cool the data centers that store and process massive datasets. This intensive energy use contributes to global warming and can have broader social, health, and safety implications [24]. As the environmental impact of these technologies continues to rise, concerns about their carbon footprint have motivated the emergence of Green AI paradigm. This paradigm aims to mitigate these impacts by optimizing algorithms, improving hardware efficiency, and adopting sustainable data-management strategies. Approaches under the Green AI umbrella emphasize low-carbon computation, smaller and more efficient models, reduced complexity, and greater transparency. These practices support energy-efficient solutions across cloud infrastructures as well as mobile and edge devices [20].

Key areas of Green-AI research include: (i) Minimizing energy usage: Creating AI models and algorithms that demand less energy during training and operation; (ii) Harnessing renewable energy sources: Utilizing clean energy options like solar and wind power to drive AI processes; (iii) Optimizing hardware: Engineering AI-specific hardware designed for superior energy efficiency [132].

A related distinction in the literature contrasts **Red AI** and Green AI [133]. According to them, Red AI refers to research that prioritizes performance gains such as higher accuracy by scaling up computational resources without regard for cost or environmental impact. In contrast, Green AI emphasizes achieving strong results while explicitly accounting for computational efficiency and resource usage. While Red AI has driven rapid increases in carbon-intensive computation, Green AI encourages methods that balance performance with sustainability, promoting more favorable performance to efficiency trade-offs.

Strategies for Green AI

Green AI is not limited to a philosophical shift toward sustainability; it also encompasses a set of concrete technical strategies designed to reduce computational cost while maintaining

competitive model performance. These strategies operate at different levels of the machine learning pipeline, including model architecture, training procedures, and data management. This subsection outlines the foundational model-level techniques that enable efficient and environmentally responsible AI development.

Model-Level Strategies

Model-level Green AI techniques focus on reducing the size, complexity, and computational demands of machine learning models. These approaches are particularly relevant for modern NLP systems, where transformer-based architectures often contain hundreds of millions of parameters and require substantial energy for both training and inference.

Pruning. Pruning is the process of eliminating unnecessary characteristics from a model to make it faster and lighter without sacrificing performance. Pruning deep neural networks can minimize model size and computational complexity without substantially affecting accuracy [98]. It can be done in a number of ways, including Weight Pruning, Magnitude-Based Pruning, Unit Pruning and Filter Pruning.

Quantization. A set of techniques known as quantization techniques is employed in deep neural networks to decrease numerical representation precision without sacrificing performance. In order to minimize memory use and computational costs during inference, these methods seek to encode weights and activations using fewer bits than conventional floating-point representations [18]. Weight quantization, activation quantization, and hybrid quantization methods are a few of the quantization techniques available; each has a unique strategy for maximizing model efficiency.

Knowledge Distillation. Transferring knowledge from a big, complicated model (teacher model) to a smaller, simpler model (student model) is known as knowledge distillation. With less constraints, the student model learns to imitate the teacher's actions. This method is very helpful for implementing AI systems on gadgets with limited processing power, like smartphones or Internet of Things devices [35].

Lightweight Transformer Architectures. Efficiency has been specifically considered in a number of transformer variations. To lower the number of parameters and computational expense, models employ strategies like factorized embeddings, bottleneck layers, or efficient pretraining objectives. These lightweight architectures provide strong baselines and offer a sustainable alternative to full-scale models.

Training-Level Strategies

Training-level Green AI strategies focus on reducing the computational cost associated with model optimization. These techniques aim to minimize the number of trainable parameters, shorten training duration, and avoid unnecessary computation, all while maintaining strong task performance.

Parameter-Efficient Fine-Tuning (PEFT). PEFT proposed by [60], involves freezing most of the parameters of pre-trained language models while updating a limited number of task-specific parameters. This saves or transfers just one general pre-trained language model along with the adjusted parameters for each task. PEFT equals the performance of full fine-tuning with less than 1% of the PLM parameters updated, aside from memory and training cost savings [87].

Early Stopping. Early stopping is a training strategy that halts the learning process as soon as the performance of the model on validation set stops improving. This prevents overfitting while avoiding unnecessary training epochs, thereby reducing energy consumption. As a simple yet effective strategy, early stopping is widely used in practice and aligns naturally with Green AI principles.

Data-Level Strategies

The goal of Data-Level techniques is to lessen reliance on big annotated datasets, which are frequently expensive to produce and manage. These methods increase data efficiency and allow models to learn efficiently with less guidance.

Weak Supervision. Weak supervision refers to training models using automatically generated or noisy labels instead of fully manual annotations. It typically relies on heuristics, rules, or external knowledge bases to create large amounts of approximate training data.

Active Learning. The basic concept of active learning is that if a machine learning algorithm is given the freedom to select the data it learns from, it can perform better with less training. In many recent machine learning and data mining applications, where unlabeled data may be plentiful or readily available but training labels are costly, time-consuming, or difficult to get, this kind of technique is highly motivated [134].

Self-Supervision. Self-supervised learning (SSL) is a machine learning paradigm in which a model is trained on a task utilizing the data itself to produce supervision instead of depend-

ing on labels from outside. SSL begins with pseudo-label training to initialize the model, then supervised or unsupervised learning is used to complete the task.

2.3 Multimodal AI

A modality describes the form through which information is conveyed or experienced. In everyday contexts, the term is often linked to sensory channels such as vision, hearing, or touch, which serve as our primary means of perceiving and communicating. A task or dataset is therefore considered multimodal when it incorporates information from more than one of these channels [13].

In artificial intelligence, multimodal understanding refers to the ability of models to process and combine information from different modalities such as text, audio, and images, to achieve a more complete interpretation of the input. For example, summarizing a soccer match requires integrating visual information (player actions, goals) with audio cues (crowd reactions, commentary) [53]. Growing interest in this area reflects the recognition that multimodal data enables AI systems to form richer and more accurate representations of complex events. Despite its potential, multimodal learning presents significant challenges. One of the most difficult aspects is effectively integrating multiple modalities so that their complementary information is captured. Achieving this requires careful architectural design, appropriate training strategies, and a solid understanding of how different modalities interact and influence one another.

Within this broader context, **Memes** have emerged as a notable medium for communicating these concepts in an engaging manner [34]. Memes are ubiquitous form of multimedia that are created by overlaying text onto images. These humorous or satirical messages have gained immense popularity as a means of communication, spreading rapidly among individuals. Although the majority of internet memes are harmless and amusing, some of them are the source of spreading inappropriate or hateful content. Manually monitoring such material is extremely difficult due to the sheer volume of data circulating online.

Figure 2.1: Examples of multimodal memes where the image and text appear harmless when viewed separately but convey a hateful meaning when combined. Source: MAMI Dataset [48].

Automated detection methods face additional challenges because memes are inherently multimodal: understanding them requires interpreting both the visual component and the accompanying text, as well as the contextual relationship between the two. While humans possess an inherent capability to comprehend the meaning conveyed by the fusion of text and images in memes, machines struggle to perform this type of complex task. In many cases, neither the image nor the text alone carries an explicitly hateful message; it is their combination that produces a harmful or derogatory meaning. Figure 2.1 provides an example of this phenomenon where the textual caption and the visual content seem harmless when viewed separately, but when combined, they suggest hatred. This reliance on the interaction between text and image makes hateful memes especially difficult for automated systems. They must not only understand each modality individually but also interpret the hidden meaning that appears when the two are combined. Therefore, multimodal hate speech detection requires models that can capture contextual associations, subtle semantic clues, and culturally grounded interpretations that people can understand clearly.

2.4 Contrastive Learning

The roots of contrastive learning trace back to the 1990s, when the idea was first introduced alongside the "Siamese" network architecture [22]. The general idea of contrastive learning is to choose an anchor sample, push it away from 'negative' samples in the embedding space, and pull it toward 'positive' examples. In unsupervised settings, a positive pair usually consists of data augmentations of the anchor, while negative pairs consist of the anchor and randomly chosen samples. A contrastive loss function such as max-margin, triplet or N-pairs loss, is used as a training objective.

Khosla et al. [71] extended this approach to the supervised setting by analyzing the two versions of new supervised loss functions that allow multiple positives per anchor. This supervised extension addresses a key limitation of the unsupervised approach: the inability to distinguish between semantically similar samples that lack explicit label information. In unsupervised contrastive learning, such samples may be incorrectly treated as negatives, leading to suboptimal representations. By incorporating label supervision, the supervised contrastive framework allows multiple semantically related samples to be treated as positives, thereby improving the model's ability to learn more discriminative and meaningful embeddings. As shown in Figure 2.2, the supervised setup correctly identifies both visually similar and semantically related samples as positives, whereas the unsupervised variant misclassifies some of them as negatives due to the absence of label guidance. This refinement has proven espe-

cially beneficial in tasks where fine-grained semantic distinctions are crucial, such as image classification, natural language understanding, and multimodal learning.

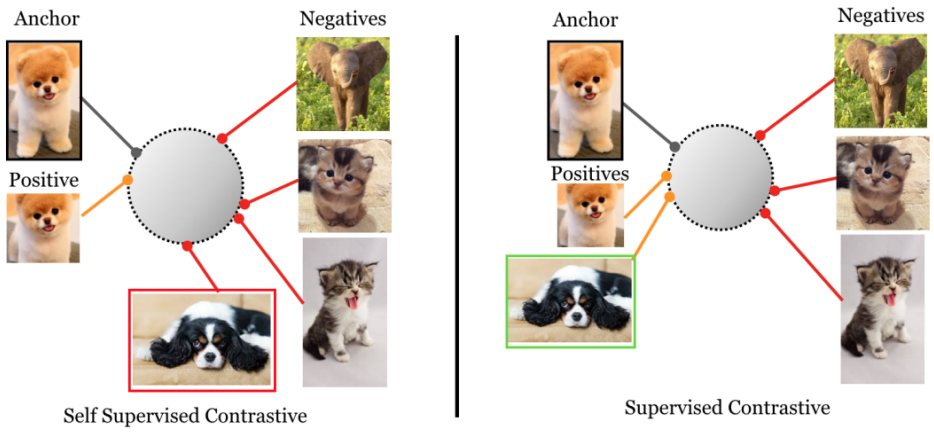


Figure 2.2: Illustration of the difference between unsupervised contrastive learning and the supervised contrastive learning framework. Source:[71]

3 Literature Review

This chapter reviews the key research areas that form the foundation of this thesis. It begins with a review of the research work on hate speech, misogyny, and sexism, outlining the challenges involved in detecting harmful content online. It also examines methodological advances such as contrastive learning and semi-supervised approaches, which offer effective solutions when labeled data is limited. Then a review of research on disagreement is included. Subsequently, the chapter discusses recent work in Green AI, emphasizing the growing importance of efficiency and sustainability in modern machine learning. Finally, it presents the state-of-the-art in vision–language models, providing an up-to-date perspective on the multimodal architectures. Together, these areas provide the background needed to understand the contributions of this thesis.

3.1 Hate Speech

The detection of hateful and abusive content on social media has attracted significant research attention, beginning with early text-focused approaches and progressively expanding toward multimodal and semi-supervised techniques. Machine learning methods including Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression were primarily used by early detection systems. To represent text data, these approaches frequently rely on handcrafted features such as Term Frequency-Inverse Document Frequency (TF-IDF) [33, 3], Bag of Words (BoW) [97] or n-grams [151] while a few others [120, 54] focus on investigating the phenomenon from a linguistic point of view. Although these methods have been proven effective, they usually require extensive feature engineering and had difficulty capturing deeper semantic cues and contextual subtleties.

The adoption of deep learning methods [37, 15] that use neural networks for automatic feature extraction have become more popular in recent years. Methods like Long Short Term Memory (LSTM) networks and Convolutiona Neural Networks (CNNs) have been used to improve detection accuracy, because they are able to identify the complex patterns in data [103, 135, 70, 4]. These methods substantially improved representation quality and classification accuracy for textual data. Multi-task learning has also been explored to transfer

knowledge across related tasks, with [68] showing that leveraging multiple classification objectives can enhance abusive language detection. Similarly, phonetic-aware deep models have been applied to Vietnamese social networks, where [79] demonstrated that capturing syllable structures improves diacritics generation and downstream hate speech detection.

Beyond text, researchers have explored contextual and structural signals. [100] proposed DRAGNET++, which models tweet reply chains to predict hatred intensity, while [102] used hypergraph convolution to capture fine-grained features. To improve detection in low-resource settings, [138] introduced the Knowledge Augmented Abusive Language Detection (KALD) framework, integrating contextual reconstruction and contrastive learning.

With the advent of transformer-based architectures, large language models have been increasingly employed for the detection of hateful content [147, 29]. These models use extensive pretraining on a variety of corpora and deep contextual knowledge to detect implicit kinds of hate speech that are frequently overlooked by conventional approaches.

Hateful content on social media often appears in multimodal formats, such as memes that combine text and images, where text-only approaches fall short. To address this, recent studies have shifted toward multimodal approaches, from simple fusion techniques [47] to sophisticated vision-language models that jointly learn text–image representations, such as VisualBERT [84], CLIP [117], and BLIP [81]. These models have shown good performance in detecting hate and offensive memes. They are frequently improved using attention-based architectures [10] or advanced fusion methods [155]. In this context, [146] highlighted the importance of cross-modal consistency, showing that visual and textual modalities may be discordant and proposing a CLIP-based architecture to identify the modality responsible for misclassification.

3.2 Sexism and Misogyny

Sexism and Misogyny is one of the crucial aspects of abusive speech online. Given its damaging impact on social media platforms, the detection and mitigation of sexist and misogynistic content has become an important area of study. [8] made an influential contribution by introducing the first dataset of misogynistic tweets and utilizing machine learning models for classification. As the field evolved, researchers realized the fact that hate speech spreads not only through text but also through visual elements.

Memes are currently one of the most popular multimodal means of communication on social media which typically combine an image with a text. Many studies have implemented both unimodal and multimodal techniques to analyze them. The first contribution in this area was by [47], who investigated the unimodal and multimodal approaches on the Memes

dataset to detect misogynous content. Another main contribution in addressing the problem of automatic detection of misogyny in memes was presented by the SemEval2022 Task 5 [48] in which the tasks on misogynous content identification and categorization were presented as a challenge.

Researchers have used transformer and non-transformer methods to detect misogyny and sexim in memes. [46] employed convolutional neural networks (CNNs) with traditional word tokenization techniques to detect sexism in augmented dataset. [157] used Bi-LSTM with transfer learning to differentiate harmless samples from sexism and other type of hateful content.

Following this, [125] evaluated four unimodal and three multimodal strategies to assess which modality contributes more effectively to misogynous meme detection. Their work also introduced a bias estimation technique to identify potentially unfair elements and a bias mitigation strategy to remove them. Similarly, [76] also proposed novel metrics to quantify model bias and evaluated strategies to mitigate this bias. [154] proposed a brain-inspired multimodal fusion framework that enhances hateful meme detection by progressively reinforcing auxiliary contexts onto main semantics to mitigate cognitive biases. [14] investigates an attention-based approach to improve performance over baseline multimodal models, like CLIP and ViLT, by assigning separate importance to the textual and visual representations. [108] rely on the CLIP model to extract the embedded text and image, and then combine them by diagonal multiplication to obtain the classification models.

Another line of work focuses on representation-learning approaches, which typically rely on lightweight architectures. Examples include MOMENTA [115], Prompt-Hate [25], and HateClipper [74]. MOMENTA enhances intra-modality attention by incorporating external facial recognition features and background knowledge into the CLIP framework. PromptHate first converts images into textual descriptions and then performs classification using a language model. HateClipper generates a text-image interaction matrix for the integration of multimodal information. Overall, these methods provide efficient classification pipelines while requiring fewer parameters.

In a recent work by [122], the researchers proposed a multimodal framework for detecting misogynistic content by adaptively fusing visual-textual features through attention mechanisms, graph-based refinement, and content-specific feature learning. The approach incorporates misogyny lexicons and test-time augmentation to improve generalization in identifying toxic patterns. Several studies have since explored misogyny detection across multiple languages beyond English [107, 120, 136]. More recent shared tasks such as IberLEF [28] and EXIST [112], now explicitly address multilingual misogyny detection and continue to engage a growing international research community.

3.2.1 Contrastive Learning based Approaches

With the advent of contrastive learning, researchers have applied contrastive learning based solutions for a variety of problems [123, 85]. In the context of unimodal sexism and misogyny detection, several studies have explored contrastive learning to enhance representation quality. For example, [92] jointly optimize self-supervised and supervised contrastive losses to capture span-level information that goes beyond token-level emotional semantics. Similarly, [121] incorporate a threshold-based contrastive learning strategy in which cosine similarities between embeddings are computed, and only sample pairs exceeding a learnable similarity threshold are treated as positive pairs.

The CLIP model developed by OpenAI [117] has shown to be a creative solution for many multimodal challenges. [88] developed a multimodal framework for the detection of hateful memes using Hateful memes dataset by Facebook AI [72]. In their study, they upsampled contrastive samples to promote multimodality and employed cross-validation ensemble learning to increase robustness, hence achieving better performance than current multimodal techniques.

Additionally, [30] experimented with contrastive learning to detect misogynous memes by using language content from the meme followed by the appropriate label as the training text for CLIP model. Furthermore, they carried out an exploratory study of the training data to determine which features are more important for a class. Other study introduced knowledge-augmented frameworks that integrate contextual and category-based contrastive learning [138]

RMIT-IR [137] proposed two supervised contrastive learning models based on CLIP (i.e., Text-Image multimodal model via CLIP-guided learning (TI-CLIP) and Text-Image Multi-View multi-modal model via CLIP-guided learning (TIMV-CLIP)). In a recent study by [159], researchers used Semantic Contrastive Learning based solution to detect toxic euphemisms on social media. They also employed a dual-channel module to enrich toxic comments with background knowledge, enhancing the detection of toxic euphemisms.

In another recent study,[140] introduces a contrastive learning framework called HATE-SIEVE that addresses the lack of fine-grained meme annotations. Its Contrastive Meme Generator (CMGen) constructs contextually aligned triplet datasets by generating semantically similar but contrasting hateful and non-hateful memes. Combined with an Image–Text Alignment module, this enables the model to learn subtle distinctions in multimodal content and improves segmentation and classification performance.

3.2.2 Semi-Supervision based Approaches

Although deep learning and transformer-based models have achieved strong performance in misogyny and sexism detection, they are highly data-dependent. Large-scale annotated datasets are costly and time-consuming to obtain, especially for low-resource languages. [106] describe this challenge as *data voracity*, highlighting the difficulty of scaling supervised models without significant annotation efforts. Furthermore, supervised methods often struggle to generalize across domains, platforms, and modalities, where expressions of hate vary considerably.

To reduce reliance on exhaustive labeled data, researchers have explored semi-supervised learning (SSL), which leverages small labeled datasets alongside large pools of unlabeled data. Pseudo-labeling has been widely adopted in this context, though its effectiveness depends heavily on data characteristics and selection strategies [96]. In this direction, several unimodal clustering-based approaches have been proposed for sexism and misogyny detection. [6] introduced a self-training method that iteratively labels an unlabeled corpus of five million tweets, resulting in the largest supervised Arabic OHS dataset. [11] developed a probabilistic clustering model to address the limitations of binary classifiers in capturing overlapping emotional signals in hate speech. Similarly, [41] evaluated label-propagation-based semi-supervised learning, comparing pretrained and task specific representations derived from a small labeled corpus. [129] proposed a novel framework for detecting and categorizing sexism on social media by integrating unsupervised task adaptation, semi-supervised learning, and semantically enriched representations. [2] presented semi-supervised techniques to augment the labeled dataset for multi-label, multi-class sexism classification.

Addressing multilingual challenges, [101] proposed a semi-supervised framework combining Generative Adversarial Networks (GANs) with pretrained language models such as mBERT and XLM-RoBERTa, achieving strong performance in cross-lingual hate speech detection with limited supervision. Other approaches include semi-meta-supervised techniques [116] and multi-task learning methods [68, 1], which utilize information shared across multiple related classification tasks to enhance the performance of each individual task or to improve the categorization of sexism.

For multimodal misogyny detection, a recent study [69] evaluates several semi-supervised learning (SSL) approaches on the memes dataset. The authors train a GRU-based model with GloVe embeddings and compare three SSL techniques:self-training, co-training, and mean teacher. Their results show that the self-training strategy achieves the strongest performance among the evaluated SSL methods.

3.3 Disagreement

The detection of hate speech is a complex and evolving field, largely due to its subjective nature, which often leads to disagreement during dataset annotation. Annotation disagreement occurs when different annotators assign different labels to the same data due to differences in their perspectives, cultural backgrounds, or interpretations. In hate speech detection, this is common because what one person finds offensive or hateful might seem acceptable to another. These differences come from factors like language, societal norms, and personal biases, making labeling a subjective and challenging task.

To address this challenge, the **Learning With Disagreement (LeWiDi)** paradigm was introduced [144], shifting the focus away from traditional majority voting towards models capable of learning from diverse perspectives. Rather than discarding disagreements, this approach preserves all interpretations provided by annotators. In the challenges proposed related to LeWiDi [144, 80], the objective was to establish a unified framework for learning from disagreements in both Natural Language Processing (NLP) and Computer Vision (CV). In EXIST 2023 [113], another challenge was presented to identify and classify sexism following the LeWiDi paradigm during the labeling process. It leverages the multiple perspectives, increasing dataset richness and improving the ability of models to generalize across different interpretations of harmful content. In EXIST 2024 [114][112], they extended this approach for multimodal data in the form of memes, which was another big step toward considering perspectives of annotators. These challenges aimed to provide datasets that highlight disagreements in language interpretation and image classification tasks. A related contribution is presented by [32], who also move beyond majority-vote aggregation by modeling each annotator's perspective. Their multi-annotator, multi-task framework treats the prediction of each annotator's judgment as an individual subtask while sharing a common underlying representation.

In a study by [130], the authors investigated how humans annotate data and why such disagreement occurs. They used a mixed-method approach to develop and test a multidimensional scale that captures individual differences (e.g., age, personality) influencing the decisions of annotators. Another work [126] identified a set of key properties that the metrics should have to do a fair comparison between models under LeWiDi.

Rizzi et al. [127] developed an unsupervised probabilistic framework to identify textual and visual triggers of annotator disagreement in misogynous memes. Their method employs an Element Disagreement Score (EDS) to evaluate how individual components relate to subjective labeling, utilizing mBERT embeddings to generalize for unseen terms. By aggregating these into a Multimodal Disagreement Score (MDS), the researchers demonstrated that com-

bining modalities is essential for capturing nuanced intent, matching state-of-the-art performance with significantly fewer parameters.

In multilingual context, [56] developed a multilingual XLM-R transformer model fine-tuned on English, Italian, and Slovenian datasets to explore how inter-annotator disagreement can be effectively incorporated into the training pipeline Their methodology compared several strategies, such as Duplicate All (DA) which retains all annotations and Duplicate Disagreement (DD), while employing a weighted cross-entropy loss function to address class imbalance in minority hate speech categories. In the context of Turkish social media, Dehghan et al. [36] fine-tuned a BERTurk transformer model to evaluate various strategies for managing annotator disagreement in multi-label hate speech classification and strength prediction. Their methodology compared several aggregation techniques to establish ground truth from subjective labels, including weighted majority voting and specific strategies such as selecting the minimum, maximum, or mean label when no clear majority existed.

One recent study [128] proposes a lightweight approach that models perspectivism by combining contextualized embeddings of sentence components with weighted probability functions. Their findings show that only a small set of linguistic and contextual cues can effectively act as proxies for identifying sentences likely to be perceived differently across annotators. Notably, this method achieves strong performance without relying on large or computationally expensive language models, underscoring the value of efficient, interpretable representations in perspectivism-aware hate speech detection.

The study of literature shows that no existing research has focused on the detection of annotator disagreement in multimodal datasets, a gap that this work addresses for the first time paving the way for the development of more robust and reliable models for disagreement detection.

3.4 Green AI

Despite advances in supervised learning approaches for hate speech detection [156, 122], their success largely depends on the availability of large-scale annotated datasets. One of the central challenges in this domain is data voracity, reflecting the substantial demand for high-quality labeled data that can be costly and difficult to procure. To overcome this limitation, researchers often rely on pretrained or foundation models trained on other tasks and transfer them to hate speech detection [106]. However, such transfer may not always capture the nuances of hate speech, especially in underrepresented languages and multimodal settings. Moreover, the reliance on large-scale pretrained models introduces significant computational demands, with studies showing that fine-tuning and deployment of such architectures can

incur substantial energy usage and carbon emissions [95]. For instance, training and fine-tuning large Transformer architectures has been estimated to produce emissions equivalent to the lifetime emissions of five US cars [139]. More broadly, the rapid growth of AI pipelines is projected to account for nearly 2% of global carbon emissions by 2030 [94], underscoring the urgent need for sustainable practices in machine learning research. Moreover, the Stanford report summarized in Figure 3.1, shows the carbon emissions associated with training frontier AI models over recent years. Their estimation tool computes emissions based on several factors, including the type of hardware used, total training hours, cloud provider, and geographical training region. According to the report of Stanford AI (illustrated in the Figure 3.1), the carbon footprint of training state–of–the–art models has increased steadily, reflecting the escalating computational demands of modern AI systems.

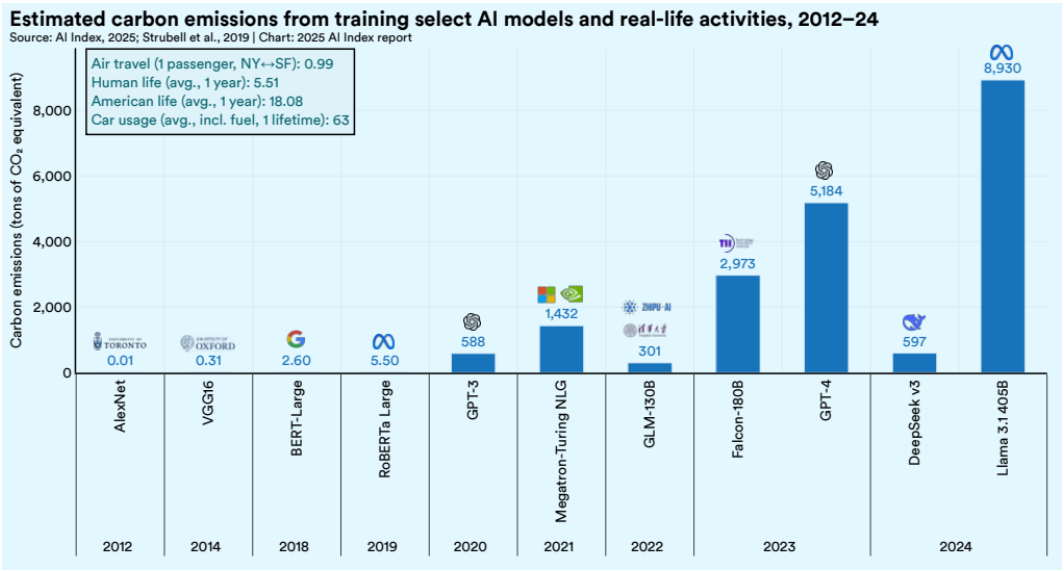


Figure 3.1: Stanford AI Index report illustrating the increase in carbon emissions from training frontier AI models over recent years.

Due to the growing climate crisis, a growing body of research is focusing on reducing the environmental impact of Deep Neural Networks (DNNs) [132]. Green AI techniques cover a broad range of strategies, including model-level compression, effective training processes, and data-centric methods, all of which work together to reduce computing cost without sacrificing model performance.

At the model level, pruning and quantization remain among the most widely explored techniques. [19] examined pruning and quantization strategies designed for NLP architec-

tures, demonstrating that substantial reductions in model size can be achieved with minimal performance degradation. Similarly, [109] introduced pruning and quantization algorithms specifically designed for Vision Transformers, enabling their deployment on resource-constrained edge devices. Beyond compression, lightweight transformer architectures have also emerged as a promising direction. ALBERT [77], a lighter version of BERT model, significantly reduces parameter count through extensive parameter sharing, while F-ALBERT [73] further lowers computational cost by combining distillation with progressively reduced parameter reuse.

Training-level methods complement these architectural innovations by reducing the cost of fine-tuning large models. Parameter-efficient fine-tuning (PEFT) approaches such as LoRA [61] introduce small, trainable modules into transformer layers while keeping the majority of parameters frozen, drastically lowering memory requirements. QLoRA [38] extends this idea by combining PEFT with quantization, enabling the fine-tuning of extremely large models (up to 65B parameters) on a single GPU without sacrificing performance. More recently, GreenTrainer [64] proposed an adaptive fine-tuning framework that selectively updates tensors based on their computational cost and contribution to accuracy, optimizing training according to FLOPs reduction objectives and aligning model adaptation with Green AI principles.

Data-level strategies also play a crucial role in reducing computational demand. Techniques such as weak supervision, active learning, and self-supervised representation learning aim to minimize reliance on large annotated datasets, which are often expensive to produce and require repeated training cycles. For instance, [132] introduced the "Play it Straight" approach, which integrates efficient data selection with incremental training inspired by active learning. By prioritizing informative samples and avoiding redundant computation, such methods reduce both labeling effort and training cost.

Another study introduces a weakly supervised deep learning framework aimed at identifying users involved in hateful interactions [124]. The approach not only provides a quantitative assessment of individuals who frequently engage in such behavior but also offers a qualitative examination of indirect or subtle hateful exchanges. Unlike models that operate at the level of individual posts or users, this method evaluates interactions, enabling a more nuanced understanding of how hateful conversations unfold and who participates in them.

Together, these model-level, training-level, and data-level innovations illustrate the breadth of current efforts to make machine learning more sustainable. The state of the art demonstrates that substantial efficiency gains are achievable through compression, selective fine-tuning, and data-efficient learning, providing a strong foundation for developing environmentally responsible systems across domains including hate speech detection.

Despite the growing emphasis on sustainability in machine learning, the application of Green AI principles to hate speech detection remains limited. Most existing systems rely heavily on pretrained vision language models such as CLIP, BLIP or GPT-based architectures. While these models achieve strong performance, their training and fine-tuning require substantial computational resources, making them costly to deploy at scale and environmentally burdensome.

A key challenge in hate speech detection, particularly for misogyny and sexism, is the lack of high-quality labeled datasets. This scarcity often motivates the use of large pretrained models, but such dependence increases energy consumption and carbon emissions.

The lack of Green AI practices in hate speech detection is especially problematic for underrepresented languages and resource-constrained settings, where computational infrastructure may be limited. As a result, there is a pressing need for models that balance accuracy with efficiency, enabling scalable and environmentally responsible content moderation. This thesis addresses this gap by evaluating supervised and unsupervised methods through the lens of Green AI, emphasizing both predictive performance and computational efficiency.

3.5 State-of-the-Art Vision–Language Models

The increasing prevalence of multimodal content on social media has accelerated research on models capable of jointly interpreting visual and textual information. Early multimodal architectures, such as **VisualBERT** [83], **ViLBERT**[91], **LXMERT**[142], and **UNITER** [27], laid the foundation for vision–language understanding by extending transformer-based architectures to process images and text in parallel. These models typically rely on region-based visual features extracted from object detectors (e.g., Faster R-CNN), which are then fused with textual embeddings through cross-modal attention mechanisms. Their success demonstrated the value of deep cross-modal alignment for tasks such as visual question answering, image–text retrieval, and multimodal classification.

Subsequent models introduced more efficient and scalable pretraining strategies. Approaches such as **OSCAR** [86], **VinVL** [158], and **ALBEF** [82] improved alignment by incorporating object tags, enhanced visual encoders, or contrastive objectives that better capture fine-grained semantic relationships. These models enabled stronger generalization across diverse multimodal benchmarks and reduced reliance on heavy region-based pipelines.

The advent of large-scale contrastive pretraining, such as **CLIP** [117], marked a significant change. It learns a shared embedding space where images and textual descriptions are aligned using a contrastive objective after being trained on hundreds of millions of image–text pairs. This paradigm unlocked powerful zero-shot and few-shot capabilities, making CLIP a widely

adopted backbone for downstream multimodal tasks, including hateful content detection and meme analysis.

Generative multimodal models further expanded the landscape. Architectures such as **BLIP** [81] combine contrastive, captioning, and image–text matching objectives to produce richer and more flexible representations. These models support both discriminative and generative tasks, offering strong performance on retrieval, captioning, and multimodal reasoning.

More recently, extremely large multimodal systems such as **Flamingo** [5], **PaLI-X** [26], **GIT** [150], **Kosmos-2** [111], and other Large Vision–Language Models (LVLMs) have demonstrated impressive capabilities in open-ended multimodal reasoning, instruction following, and few-shot learning. These models integrate massive image–text corpora, sophisticated fusion mechanisms, and large-scale language models to achieve state-of-the-art results across a wide range of benchmarks. However, their substantial computational requirements, high energy consumption, and large carbon footprint limit their practicality for real-time or resource constrained applications such as content moderation.

Within this diverse ecosystem, models like CLIP and BLIP represent a balanced middle ground: they offer strong multimodal alignment, broad applicability, and competitive performance while remaining significantly more lightweight than the largest LVLMs. Their efficiency, accessibility, and widespread adoption make them suitable baselines for evaluating multimodal classification tasks, including misogyny detection, hate–disagreement modeling, and environmentally conscious experimentation.

4 Datasets and Evaluation Methods

This chapter presents the datasets and evaluation methods used throughout this thesis. It first describes the MAMI and EXIST datasets, outlining their structure, annotation schemes, and relevance to the tasks addressed in this work. The chapter then introduces the vision-language models used as baselines. Following this, the chapter outlines the related measures for assessing model performance and describes the carbon-footprint estimation procedure used to quantify the environmental impact of the experiments. Together, these components establish the basis for the methodologies presented in the subsequent chapters.

4.1 Datasets

We utilized two publicly available benchmark datasets to evaluate our proposed methods for detecting misogyny, sexism and disagreement in memes. These datasets are specifically designed for multimodal social media content analysis and offer rich annotations relevant to our task. We provide detailed descriptions of both datasets below.

4.1.1 MAMI

We used the benchmark dataset for the **Multimedia Automatic Misogyny Identification (MAMI)** challenge held at SemEval 2022 [48] to assess our proposed methods. The dataset consists of memes collected from multiple online platforms and supports two related classification tasks:

Task A: Misogynous Meme Identification A binary classification task where each meme is labeled as either *misogynous* or *non-misogynous*.

Task B: Misogyny Type Classification A multi-label task where misogynous memes are further categorized into potentially overlapping subtypes:

Figure 4.1: Examples of some misogynous memes from MAMI dataset

- **Shaming:** Content that insults or demeans women based on physical traits, personality, or behavior.
- **Stereotype:** Memes reinforcing fixed or conventional ideas about women, including role-based or gender-based stereotypes.
- **Objectification:** Depictions that treat women as objects rather than individuals.
- **Violence:** Content that encourages or depicts physical harm or threats against women.

Figure 4.1 shows some examples of misogynous memes from MAMI dataset.

Dataset Statistics

The dataset consists of 10,000 training and 1,000 testing memes. Table 4.1 describes the details of MAMI dataset.

Task	Label	Train	Test
Dataset size		*10,000*	*1,000*
Misogyny	Misogynous	5,000	500
	Non-misogynous	5,000	500
Agreement	Agreement	6,208	–
	Disagreement	3,492	–

Table 4.1: Statistics of the MAMI dataset. Both tasks are applied to the same 10,000 training and 1,000 test memes. Agreement labels are only available for the training split.

Data Collection Procedure

The dataset was constructed by gathering misogynous and non-misogynous memes from a variety of online sources. The collection process involved:

1. Searching major social media platforms such as Twitter and Reddit.
2. Scraping and manually downloading memes from popular meme-sharing websites including 9GAG, KnowYourMeme, and Imgur.

Annotation Protocol and Disagreement Labels

Each meme was independently annotated by three annotators. The final binary misogyny label was determined through majority voting: if at least two annotators agreed on a label, that label was assigned as the ground truth.

Using this annotation information, we derive labels for disagreement in our work:

- A sample is labeled **agreement (0)** if all three annotators assigned the same label.
- A sample is labeled **disagreement (1)** if there is any inconsistency in the labels, indicating a lack of consensus.

The disagreement labels are only available for the Training set (10,000) memes. These labels are used in our work to study annotation uncertainty and model robustness.

4.1.2 EXIST

The second dataset used in our study comes from **EXIST 2024, the fourth edition of the sEXism Identification in Social neTworks challenge** [112, 114]. The challenge includes several tasks related to sexism detection in textual and multimodal content, including memes. Specifically, the meme-related tasks are defined as follows:

- **Sexism Identification in Memes:** A binary classification task aimed at determining whether a meme is sexist or non-sexist.
- **Source Intention in Memes:** A task focused on identifying the author's intention behind the meme, with two possible classes: DIRECT or JUDGEMENTAL.
- **Sexism Categorization in Memes:** A multi-class task that categorizes sexist memes into one of the following five categories: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence, and (v) misogyny and non-sexual violence.

Figure 4.2: Examples of some sexist memes from EXIST dataset

In our work, we selected the binary classification task (Sexism Identification in Memes) to detect sexist versus non-sexist memes. Figure 4.2 shows some examples of sexist memes from this dataset.

Dataset Statistics

The dataset includes memes in both English and Spanish, with 4,044 memes for training and 1,053 memes for testing. Table 4.2 summarizes the dataset statistics.

Category	Train	Test
Dataset Size		
Total Memes	4,044	1,053
Spanish	2,034	540
English	2,010	513
Sexism Annotation		
Sexist	2,038	–
Non-sexist	1,382	–
Unknown	624	–
Agreement Annotation		
Agreement	957	–
Disagreement	3,087	–

Table 4.2: Statistics of the EXIST dataset. The dataset contains English and Spanish memes. Labels for sexism and agreement/disagreement are available only for the training split.

Data Collection Procedure

The dataset creators first curated a lexicon of terms and expressions commonly associated with sexist content. This lexicon consists of 250 seed terms, including 112 English and 138

Spanish terms. These terms were used as search queries on Google Images to retrieve the top 100 images for each query.

A rigorous manual cleaning process was then applied to ensure data quality. The final dataset contains more than 3,000 memes per language.

Annotation Protocol and Disagreement Labels

The dataset follows the LeWiDi paradigm, providing label information from all annotators. Each meme was annotated by seven annotators, and demographic information such as ethnicity, gender, and age was also recorded.

For each meme, the individual labels were collected, and a final hard label was assigned using majority voting:

- If at least four out of seven annotators agreed on a label, that label was assigned as the ground truth.
- If the votes were evenly split (e.g., three votes for "sexist" and three for "non-sexist"), the meme was labeled as **unknown** to reflect ambiguity.

For our disagreement detection task, we derived labels based on the individual annotations. A meme was labeled as:

- **agreement (0)** if all seven annotators assigned the same label,
- **disagreement (1)** if there was any variation among the annotators' labels.

These derived labels allow us to study annotation uncertainty and model robustness.

4.2 Vision Language Models used as Baselines

Building on the state-of-the-art multimodal models discussed in Chapter 3, we select **CLIP** and **BLIP** as baseline vision–language encoders due to their strong balance between performance, generalization ability, and computational efficiency. Unlike extremely large multimodal systems such as Flamingo or PaLI-X, which require substantial resources and are impractical for environmentally conscious experimentation, CLIP and BLIP offer high-quality image–text representations while remaining lightweight enough for systematic evaluation under varying supervision levels. They are ideal for tasks including meme understanding, hateful content identification, and complex semantic interpretation due to their proven efficacy across a variety of multimodal benchmarks. Moreover, their moderate computational footprint allows for a fair and transparent comparison of performance and carbon emissions, aligning with the Green AI principles.

4.2.1 CLIP (Contrastive Language–Image Pre-training)

CLIP [117] is a multimodal model designed to learn a shared understanding of images and natural language. Its central idea is straightforward: the model learns to place an image and its correct text description close together in a common embedding space, while pushing apart images and texts that do not match. This is achieved by training on a very large collection of image–text pairs naturally found on the internet, such as captions, alt-text, and descriptions. Because this supervision is abundant and diverse, CLIP learns a broad range of visual concepts without relying on manually curated labels. CLIP consists of two separate encoders:

- An image encoder (e.g., ResNet or Vision Transformer) that converts an image into a vector representation.
- A text encoder (a transformer-based model) that converts a text description into another vector representation.

Both encoders map their inputs into the same embedding space, allowing direct comparison between images and text. During training, CLIP processes a batch of images and their corresponding text descriptions, computes embeddings for all of them, and then compares every image with every text using cosine similarity. The training objective uses the contrastive learning approach that encourages the similarity of the correct image–text pairs to be high, while the similarity of all incorrect pairings is reduced. This contrastive learning setup allows the model to associate visual content with the language used to describe it. Figure 4.3 illustrates the overall training and inference process of the CLIP model.

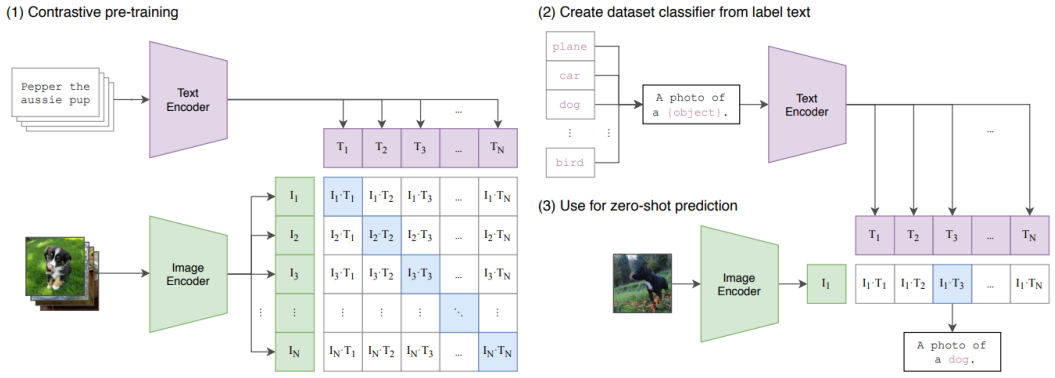


Figure 4.3: Training and inference workflow of the CLIP model. Source:[117]

A key strength of CLIP is that it can perform many classification tasks without being specifically trained for them, which is known as zero-shot classification. This flexibility arises

from the wide variety of concepts CLIP learns during training, making it applicable to many visual tasks without additional fine-tuning.

4.2.2 BLIP (Bootstrapping Language-Image Pre-training)

BLIP [81] is a vision-language pre-training framework designed for both understanding and generation tasks. It supports applications such as image captioning, visual question answering, and cross-modal retrieval by learning joint representations of images and text from large-scale web data. BLIP is particularly effective at handling noisy image–text pairs by bootstrapping and filtering captions.

BLIP introduces two main contributions. The first is the Multimodal Mixture of Encoder-Decoder (MED) architecture, which enables flexible multi-task pre-training. The MED can operate in three modes: as a unimodal encoder that processes images and text separately, as an image-grounded text encoder that integrates visual context into text representations, and as an image-grounded text decoder that generates text conditioned on images. The model is jointly pre-trained with three objectives: image–text contrastive learning, image–text matching, and image-conditioned language modeling.

The second contribution is the Captioning and Filtering (CapFilt) strategy, a data bootstrapping method for learning from noisy web image–text pairs. A pre-trained MED is fine-tuned into two components: a captioner that generates synthetic captions for web images, and a filter that removes noisy or low-quality captions from both the original and synthetic text. This procedure improves the quality of the supervision signal and enhances the effectiveness of pre-training. An overview of the BLIP framework is shown in Figure 4.4.

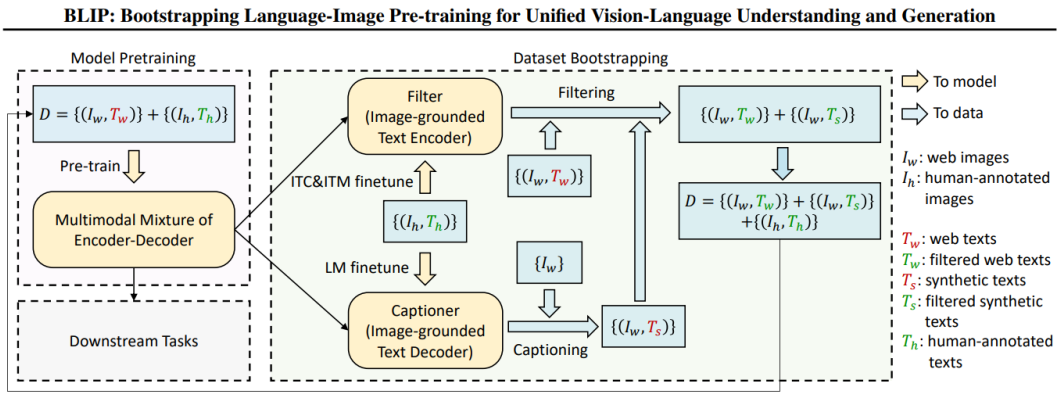


Figure 4.4: Overview of the BLIP framework, illustrating the multimodal encoder–decoder architecture and the Captioning and Filtering (CapFilt) pipeline. Source: [81]

By pre-training on millions of image–text pairs, BLIP achieves strong performance across a variety of vision-language benchmarks. Its transformer-based architecture supports rich interactions between visual and textual information, making it a versatile model for multimodal understanding and generation.

4.3 Evaluation Methods

4.3.1 Evaluation Metrics

To assess the performance of our classification models, we rely on standard evaluation metrics commonly used in binary classification: *precision*, *recall*, and the *F1-score*. These metrics are computed using four fundamental outcomes that characterize the predictions of a binary classifier:

- **True Positive (TP):** the number of positive samples correctly classified as positive.
- **True Negative (TN):** the number of negative samples correctly classified as negative.
- **False Positive (FP):** the number of negative samples incorrectly classified as positive.
- **False Negative (FN):** the number of positive samples incorrectly classified as negative.

Based on these quantities, we compute the following evaluation metrics.

Recall Recall, also known as the true positive rate (TPR), measures the proportion of actual positive samples that are correctly identified by the model. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.1}$$

Precision Precision measures the proportion of predicted positive samples that are truly positive. It evaluates how reliable the model's positive predictions are. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

F1-score The F1-score is the harmonic mean of precision and recall. It provides a balanced measure that accounts for both false positives and false negatives, making it particularly useful when the class distribution is imbalanced. It is defined as:

$$\text{F1score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

These metrics collectively provide a comprehensive understanding of model performance, capturing both correctness of positive predictions and the ability to detect positive instances.

4.3.2 Environmental Impact Analysis (Carbon Footprint)

In addition to evaluating model performance, we analyze the environmental impact of each strategy by estimating energy consumption and carbon emissions. Monitoring energy can be achieved through hardware or software tools. Hardware-based methods offer high precision [24], but they are costly and difficult to synchronize for short training runs. We therefore use a software-based approach, which is more accessible and scalable for systematic evaluation.

Following the "Green AI" methodology [95], total energy consumption (TEC) is estimated as:

$$\text{TEC} = \sum_i P_i \cdot U_i \cdot T \tag{4.4}$$

where P_i is the Thermal Design Power of the component i (in kilo watts), U_i is its average utilization (in decimal), and T is the execution time (in hours). TEC is multiplied by the carbon intensity factor ϵ (kg CO_2e/kWh) of the Italian electricity grid to obtain emissions:

$$CO_2\text{e} = \text{TEC} \times \epsilon \tag{4.5}$$

This formulation is consistent with the activity data approach [17], where emissions are derived by combining activity data (energy consumed) with emission factors specific to the region. Since our measurements rely on average utilization and regional emission factors rather than direct hardware metering, the reported values should be interpreted as *estimations* rather than exact measurements. All experiments have been conducted under identical hardware and software conditions to ensure comparability.

5 Semi-Supervised Clustering for Sexism and Misogyny Detection

This chapter investigates how pretrained vision–language models can be effectively leveraged within a semi-supervised framework to address the challenges of sexism and misogyny detection. Specifically, we introduce a constrained clustering strategy that exploits the representational power of large pretrained models without requiring substantial labeled data or expensive training cycles. By introducing a regularization mechanism directly in the embedding space, the proposed approach guides the clustering process toward semantically meaningful groupings that support misogyny and sexism detection in multimodal content.

Beyond methodological innovation, this chapter also examines the trade-off between accuracy and environmental sustainability. In line with the emerging movement toward "Green AI," we analyze how different design choices influence both model performance and energy consumption, highlighting the importance of developing systems that are not only effective but also computationally responsible.

Contributions

The main contributions of this Chapter are as follows:

1. A constrained clustering framework has been proposed that leverages pretrained vision language model embeddings to detect sexism and misogyny in multimodal content. Specifically, a regularization strategy on the embedding space has been proposed to effectively exploit the representational power of pretrained models.

2. A comprehensive evaluation on two benchmark datasets (MAMI and EXIST) using two state-of-the-art vision–language models, demonstrating competitive performance with limited supervision.

3. An analysis of trade–off between model accuracy and carbon energy consumption is performed, highlighting the efficiency of our method compared to baselines.

5.1 Methodology

We introduce a semi–supervised clustering framework for sexism and misogyny detection in multimodal content. Each meme is first represented through embeddings extracted from pre-trained vision-language models. A proportion of the dataset is labeled, while the remaining samples remain unlabeled. To incorporate supervision, we extend the classical K-Means [90] objective with a cluster-level penalty term that discourages assigning labeled points to clusters dominated by a different class. In this way, labeled samples guide the clustering process toward semantically consistent groups, while unlabeled samples are assigned based purely on distance to cluster centroids. During training, centroids are updated iteratively, and test samples are classified based on their nearest centroid. Throughout this process, we record execution time and estimate carbon emissions, enabling a joint evaluation of predictive performance and environmental impact.

5.1.1 Embedding Representation

Each meme in the datasets contains both visual and textual elements. To represent these modalities, we experimented with two vision–language models: CLIP [117] and BLIP [81]. For CLIP, the processor tokenizes the text and normalizes the images to ensure alignment, generating separate embeddings for each modality. These embeddings are then averaged to produce a single multimodal feature vector.

In parallel, BLIP is employed to extract embeddings that model fine–grained interactions between vision and language. We use the embeddings from each model independently to train our clustering algorithms.

5.1.2 Semi–Supervised Clustering

Our proposed method builds upon the classical K–Means algorithm by incorporating label information into the clustering process. The standard K–Means algorithm [90] partitions a dataset $X = \{x_1, x_2, ..., x_n\}$ into k clusters by minimizing the within–cluster variance. The original objective function is defined as:

$$J = \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|^2 \tag{5.1}$$

where μ_{z_i} is the centroid of the cluster assigned to point x_i, and $z_i \in \{1, ..., k\}$ denotes the cluster index.

Several studies have extended this formulation by introducing constraints. The first attempt in this context was by [149] to extend K-Means with *must-link* and *cannot-link* con-

straints, ensuring that certain pairs of points were either grouped together or kept apart. While effective, these pairwise constraints can be computationally expensive. In contrast, our approach introduces supervision through a cluster-level penalty rather than pairwise constraints, with the goal of being more efficient and scalable. Specifically, we extend the objective with a term that discourages assigning labeled points to clusters dominated by a different class. Let $y_i \in \{-1, 0, 1\}$ represent the label of sample x_i, where -1 indicates an unlabeled point, 0 denotes non-hate and 1 denotes hate. Our function is therefore defined as:

$$J' = \sum_{i=1}^{n} \Big(\|x_i - \mu_{z_i}\|^2 + \lambda \cdot \mathbb{I}(y_i \neq \hat{y}_{z_i}) \Big) \tag{5.2}$$

Here, λ is a penalty coefficient, $\mathbb{I}(\cdot)$ is the indicator function, and $\hat{y}_{z_i}$ is the majority label in cluster z_i. This formulation encourages labeled samples to be grouped with semantically consistent clusters, while allowing unlabeled samples to be assigned based purely on distance.

The proposed algorithm begins by initializing cluster centroids using the mean embeddings of labeled samples from each class. This ensures that the initial centroids are semantically aligned with known categories, providing a meaningful starting point for clustering. During each iteration, the algorithm processes every data point individually. For unlabeled samples, the assignment is straightforward: the point is assigned to the cluster whose centroid is closest in Euclidean space. However, for labeled samples, the assignment decision is guided not only by proximity but also by the label distribution within the candidate cluster. Specifically, if the cluster has a majority of labeled points that match the label of the current sample, the point is assigned without modification. In contrast, if the cluster's majority label differs from the sample's label, a penalty term is added to the distance calculation. This penalized distance discourages the assignment of labeled samples to semantically inconsistent clusters, thereby reinforcing label coherence within each group.

We implemented two variants of this algorithm:

- **Online Update:** Centroids are updated immediately after each individual sample is assigned to a cluster. This allows the model to incorporate new information incrementally, making the updates more responsive to recent assignments and potentially faster to adapt in streaming or large-scale settings. However, such frequent updates may introduce higher variance and reduced stability in the centroid trajectories.

- **Batch Update:** Centroids are updated only after all samples have been assigned in a given iteration. This strategy aggregates information from the entire dataset before updating the centroids, leading to more stable and consistent updates, albeit at the cost of increased memory usage and potentially slower convergence.

Algorithm 1 Batch Semi-Supervised K-Means

Initialize centroids $\mu_1, \mu_2, \ldots, \mu_k$ using means of labeled data per class
while not converged **do**
  for all points x_i **do**
    Compute distances $D_j = \|x_i - \mu_j\|^2$ for all j
    if $y_i = -1$ **then** ▷ Unlabeled point
      $z_i \leftarrow \arg\min_j D_j$
    else
      if $y_i \neq \hat{y}_{\arg\min_j D_j}$ **then**
        $D_j \leftarrow D_j + \lambda$ ▷ Penalty
      end if
      $z_i \leftarrow \arg\min_j D_j$
    end if
  end for
  Update centroids μ_j using all assigned points
  if $\forall j : \|\mu'_j - \mu_j\| < \varepsilon$ **then**
    break
  end if
end while
return $\{\mu_j\}, \{z_i\}$

Algorithm 2 Online Semi-Supervised K-Means

Initialize centroids $\mu_1, \mu_2, \ldots, \mu_k$ using means of labeled data per class
while not converged **do**
  for all points x_i **do**
    Compute distances $D_j = \|x_i - \mu_j\|^2$ for all j
    if $y_i = -1$ **then** ▷ Unlabeled point
      $z_i \leftarrow \arg\min_j D_j$
    else
      if $y_i \neq \hat{y}_{\arg\min_j D_j}$ **then**
        $D_j \leftarrow D_j + \lambda$ ▷ Penalty
      end if
      $z_i \leftarrow \arg\min_j D_j$
    end if
    Update centroid μ_{z_i} based on new assignment
  end for
  if $\forall j : \|\mu'_j - \mu_j\| < \varepsilon$ **then**
    break
  end if
end while
return $\{\mu_j\}, \{z_i\}$

These strategies allow us to explore the impact of update timing on clustering stability and performance. We present the pseudocode for our semi-supervised clustering algorithms in Algorithm 1 and Algorithm 2, which outlines the assignment and update steps for both labeled and unlabeled data.

To evaluate the robustness of our semi-supervised clustering framework, we designed experiments with increasing levels of supervision, starting from 10% labeled data and incrementally adding 10% up to full supervision. During inference, each test meme was embedded using the same vision language models and assigned to the nearest cluster based on Euclidean distance. The predicted label was determined by the majority label of that cluster, allowing us to classify unseen samples without a separate classifier.

5.2 Experimental Settings

We evaluate our proposed method using two publicly available benchmark datasets EXIST and MAMI for sexism and misogyny detection in memes (see Section 4.1). For each dataset, we used the official test sets provided to assess the generalization performance of our clustering framework.

Our experiments compared fully finetuned multimodal models (CLIP and BLIP) with semi-supervised clustering approaches, including both batch and online variants of K-Means, across varying proportions of labeled data. For each configuration, we measured model performance and carbon-footprint using the methods described in Section 4.3.

We used two models as baselines:

5.2.1 Contrastive Language—Image Pretraining (CLIP)

We adopted CLIP [117] as one of our baselines for misogyny detection in memes. CLIP is a multimodal framework trained to align visual and textual information in a shared embedding space (see Section 4.2.1 for details). In our setup, each meme was processed using the CLIP tokenizer and image preprocessor to ensure consistency and alignment between modalities. We fine–tuned the model by attaching a classification head on top of the embeddings, where image and text features were averaged to form a joint representation. This representation was then mapped to binary logits, allowing the model to predict whether a meme belongs to the misogynistic/sexist or non-misogynistic/non–sexist category. To evaluate the effect of supervision, CLIP was fine–tuned incrementally, beginning with 10% of the available training data and progressively increasing up to 100%.

5.2.2 Bootstrapped Language–Image Pretraining (BLIP)

We further employed BLIP [81] as another baseline. BLIP is a multimodal architecture designed to effectively integrate textual and visual inputs (see Section 4.2.2 for details). For these experiments, memes were preprocessed with the BLIP processor, which standardizes image features and ensures uniform text formatting through truncation and padding. We fine–tuned BLIP on our datasets for misogyny and disagreement detection by adding a classification layer that outputs binary predictions. To optimize performance while maintaining efficiency, only the top layers of BLIP were unfrozen during fine–tuning, enabling partial adaptation to our tasks while leveraging the strength of pretrained weights. Similar to CLIP, BLIP was fine–tuned with varying proportions of the training data, starting from 10% and gradually scaling up to the full dataset (100%).

5.3 Results and Analysis

This section presents a comparative analysis of model performance and environmental impact across fully fine-tuned multimodal models and semi-supervised clustering approaches. We focus on understanding the trade-offs between predictive accuracy, computational efficiency, and carbon emissions under varying proportions of labeled data. We begin by discussing results on the **MAMI dataset**, highlighting how clustering-based methods achieve competitive performance while substantially reducing computational and environmental costs.

5.3.1 MAMI

The finetuned models achieved consistently strong performance on the **MAMI dataset**, with F1-scores reaching 0.70 for CLIP embeddings and 0.66 for BLIP embeddings at 80–100% labeled data (see Tables 5.1 and 5.2). However, this came at a substantial computational and environmental cost. For example, finetuned CLIP required 73.7 minutes at 100% labeled data, resulting in 55.8 gCO_2, while finetuned BLIP consumed 94.2 minutes, producing 71.4 gCO_2. Even at moderate labeling levels, emissions remained high: at 30% labeled data, finetuned CLIP produced 15.4 gCO_2 and finetuned BLIP 22.4 gCO_2.

In contrast, clustering approaches demonstrated remarkable efficiency. Batch K-Means runs at 100% labeled data required only 7.7 minutes, yielding 2.6 gCO_2, while online K-Means consumed 9.2 minutes, producing 3.1 gCO_2. At 30% labeled data, batch clustering emitted 1.1 gCO_2 and online clustering 1.3 gCO_2, compared to the 15–22 gCO_2 of finetuned baselines. Despite this dramatic reduction in emissions, clustering achieved competitive F1-scores: for instance, at 30% labeled data, batch K-Means reached 0.66 and online K-Means 0.68,

Model	% Labeled	P	R	F1	Time (min)	gCO_2
Finetuned CLIP	10%	0.76	0.67	0.64	1.75	1.33
	20%	0.75	0.68	0.66	15.5	11.7
	30%	0.75	0.69	0.67	20.3	15.4
	40%	0.75	0.70	0.68	29.6	22.4
	50%	0.76	0.71	0.70	33.3	25.2
	60%	0.76	0.71	0.70	43.6	33.0
	70%	0.76	0.71	0.70	48.2	36.5
	80%	0.76	0.72	0.70	59.4	45.0
	90%	0.76	0.72	0.70	62.1	47.0
	100%	0.76	0.72	0.70	73.7	55.8
CLIP + K-Means (Batch)	10%	0.69	0.66	0.64	**1.59**	**0.54**
	20%	0.70	0.67	0.65	**2.52**	**0.85**
	30%	0.71	0.67	0.66	**3.15**	**1.06**
	40%	0.73	0.68	0.66	**3.84**	**1.29**
	50%	0.73	0.69	0.68	**4.51**	**1.52**
	60%	0.74	0.70	0.68	**5.19**	**1.75**
	70%	0.75	0.71	**0.70**	**6.20**	**2.09**
	80%	**0.76**	**0.72**	**0.70**	**6.57**	**2.21**
	90%	**0.76**	**0.72**	**0.70**	**7.13**	**2.40**
	100%	**0.76**	**0.72**	**0.70**	**7.77**	**2.62**
CLIP + K-Means (Online)	10%	0.69	0.66	**0.65**	**2.09**	**0.70**
	20%	0.71	0.68	**0.66**	**3.11**	**1.05**
	30%	0.74	0.70	**0.68**	**3.82**	**1.29**
	40%	0.74	0.70	0.68	**4.60**	**1.55**
	50%	0.75	0.70	0.68	**5.37**	**1.81**
	60%	0.75	0.70	0.69	**6.14**	**2.07**
	70%	0.75	0.71	0.69	**7.22**	**2.43**
	80%	**0.76**	**0.72**	**0.70**	**7.65**	**2.57**
	90%	**0.76**	**0.72**	**0.70**	**8.29**	**2.79**
	100%	**0.76**	**0.72**	**0.70**	**9.19**	**3.09**

Table 5.1: Performance comparison across label percentages for three models on **MAMI test set** using **CLIP** Embeddings, including execution time and carbon emissions. **Bold and underlined** values indicate the best performance for each metric at a given label percentage. **Bold** values indicate results that are equal to or exceed the performance of the Finetuned CLIP baseline for the corresponding label percentage.

closely matching finetuned CLIP (0.67) and BLIP (0.63). Overall, batch clustering consumed slightly less energy, while online clustering emitted marginally more CO_2 but consistently achieved better F1-scores, making it the more accurate yet still sustainable option.

Model	% Labeled	P	R	F1	Time (min)	gCO_2
Finetuned BLIP	10%	0.70	0.65	0.62	9.20	6.97
	20%	0.70	0.64	0.62	18.7	14.2
	30%	0.70	0.66	0.64	29.5	22.4
	40%	0.73	0.67	0.65	37.9	28.7
	50%	0.72	0.68	0.67	47.8	36.2
	60%	0.75	0.69	0.68	55.9	42.3
	70%	0.76	0.67	0.64	66.2	50.2
	80%	0.71	0.65	0.63	80.1	60.1
	90%	0.72	0.67	0.65	85.9	65.1
	100%	0.73	0.68	0.66	94.2	71.4
BLIP + K-Means (Batch)	10%	0.64	0.62	0.60	**1.60**	**0.54**
	20%	0.64	0.62	0.60	**2.26**	**0.76**
	30%	0.64	0.62	0.60	**2.94**	**0.99**
	40%	0.64	0.62	0.60	**3.55**	**1.19**
	50%	0.65	0.62	0.61	**4.20**	**1.41**
	60%	0.65	0.62	0.61	**4.76**	**1.60**
	70%	0.65	0.62	0.61	**5.41**	**1.82**
	80%	0.66	**0.65**	**0.63**	**5.91**	**1.99**
	90%	0.66	0.64	**0.65**	**6.55**	**2.21**
	100%	0.65	0.64	0.65	**7.19**	**2.42**
BLIP + K-Means (Online)	10%	0.65	0.64	**0.63**	**2.13**	**0.72**
	20%	0.65	0.64	**0.63**	**2.84**	**0.96**
	30%	0.64	0.64	**0.64**	**3.55**	**1.19**
	40%	0.65	0.64	0.64	**4.23**	**1.42**
	50%	0.65	0.64	0.64	**5.0**	**1.68**
	60%	0.64	0.64	0.64	**5.64**	**1.90**
	70%	0.65	0.65	**0.65**	**6.38**	**2.15**
	80%	0.66	0.64	**0.65**	**6.94**	**2.34**
	90%	0.66	0.65	**0.65**	**7.64**	**2.57**
	100%	0.66	0.65	0.65	**8.55**	**2.88**

Table 5.2: Performance comparison across label percentages for three models on **MAMI test set** using **BLIP** Embeddings, including execution time and carbon emissions. **Bold and underlined** values indicate the best performance for each metric at a given label percentage. **Bold** values indicate results that are equal to or exceed the performance of the Finetuned BLIP baseline for the corresponding label percentage.

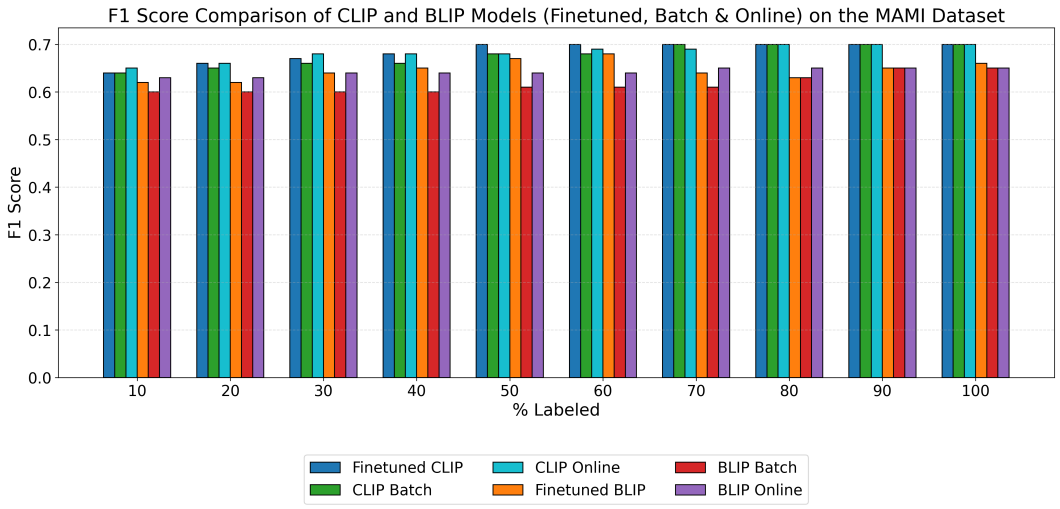


Figure 5.1: F1 scores obtained by finetuned, batch K-Means, and online K-Means variants of CLIP and BLIP embeddings across different proportions of labeled data. CLIP consistently outperforms BLIP, while clustering-based methods narrow the gap with finetuning as label availability increases.

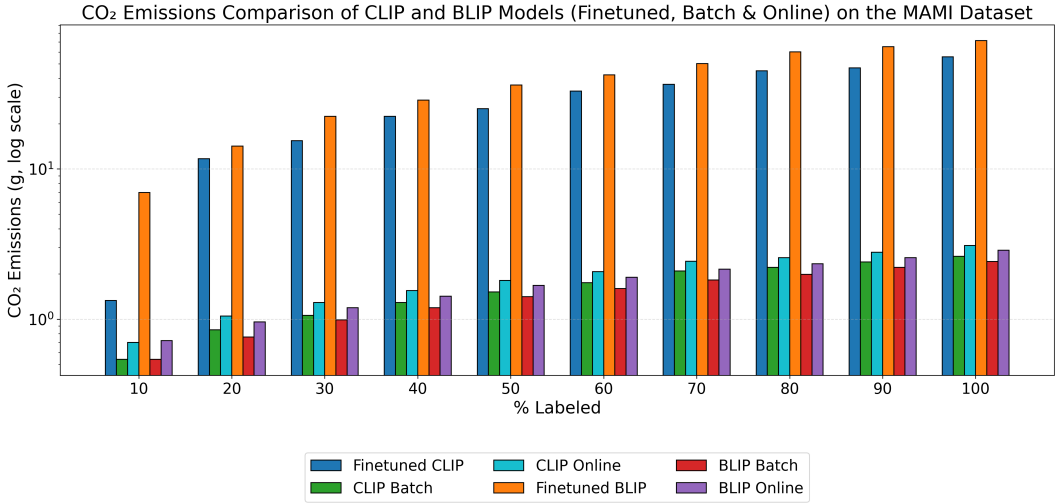


Figure 5.2: Comparison of CO_2 emissions for CLIP and BLIP models under finetuning, batch clustering, and online clustering across increasing label percentages. The logarithmic scale highlights the large environmental gap between finetuned models and the significantly more efficient clustering-based methods.

A comparison between CLIP- and BLIP-based models on the MAMI dataset (see Figures 5.1 and 5.2) shows a clearer separation between the two embedding families than what is later observed in the EXIST experiments. Across all supervision levels, CLIP consistently delivers stronger predictive performance: both finetuned CLIP and its clustering variants reach F1-scores of 0.70, whereas BLIP peaks at 0.66 even under full supervision. This advantage persists in the clustering setting, where CLIP-based batch and online K-Means frequently match or exceed the finetuned CLIP baseline while maintaining emissions between 0.5–3 gCO_2. BLIP-based clustering remains competitive but slightly weaker, typically reaching 0.63–0.65 at best. In terms of efficiency, BLIP is substantially more carbon-intensive when finetuned—emitting 7–71 gCO_2 compared to CLIP's 1–56 gCO_2—but this difference nearly disappears under clustering, where both models operate in the low-emission regime. Overall, CLIP provides a more favorable performance–efficiency profile for MAMI, with clustering narrowing the gap between the two models but not fully eliminating CLIP's advantage in predictive accuracy.

5.3.2 EXIST

Across the **EXIST English experiments** (see Tables 5.3 and 5.4), the results reveal a clear performance–efficiency trade-off between fully supervised fine-tuning and the proposed clustering based approaches. The Finetuned CLIP and Finetuned BLIP models obtain the highest F1-scores under full supervision (0.70 for CLIP and 0.64 for BLIP). However, these gains come at a substantial computational and environmental cost: training time increases almost linearly with the percentage of labeled data, reaching 47 minutes and 35.6 gCO_2 for CLIP and 22 minutes and 16 gCO_2 for BLIP at 100% supervision.

In contrast, both batch and online K-Means clustering demonstrate remarkable efficiency. Even with only 10–30% labeled data, clustering achieves F1-scores that are competitive and in several cases exceed from the finetuned models trained with the same amount of supervision. For instance, with just 10% labeled data, online clustering with CLIP embeddings reaches an F1-score of 0.60, substantially outperforming finetuned CLIP (0.43). This trend persists across label percentages: clustering maintains stable performance while keeping emissions extremely low, typically below 1 gCO_2.

A consistent pattern emerges between the two clustering variants. Online clustering tends to achieve slightly higher F1-scores, particularly at lower supervision levels, whereas batch clustering is consistently the most energy-efficient. This distinction highlights a practical trade-off: batch clustering is preferable when minimizing environmental impact is the priority, while online clustering offers a modest performance boost with only a small increase in emissions.

Model	% Labeled	P	R	F1	Time (min)	gCO_2
Finetuned CLIP	10%	0.57	0.52	0.43	5.14	3.90
	20%	0.59	0.61	0.57	9.8	7.43
	30%	0.60	0.61	0.60	14.29	10.8
	40%	0.65	0.65	0.62	19.79	14.9
	50%	0.69	0.72	0.69	24.08	18.3
	60%	0.69	0.73	0.70	28.53	21.6
	70%	0.71	0.75	0.72	33.49	25.4
	80%	0.70	0.73	0.70	38.37	29.1
	90%	0.71	0.73	0.70	42.09	31.9
	100%	0.71	0.73	0.70	46.94	35.6
CLIP + K-Means (Batch)	10%	0.54	**0.60**	**0.55**	**0.72**	**0.24**
	20%	0.55	**0.61**	0.56	**0.92**	**0.31**
	30%	0.58	**0.64**	0.59	**1.07**	**0.36**
	40%	0.59	0.64	0.60	**1.25**	**0.42**
	50%	0.60	0.65	0.61	**1.42**	**0.48**
	60%	0.62	0.68	0.63	**1.56**	**0.53**
	70%	0.62	0.68	0.64	**1.71**	**0.58**
	80%	0.63	0.70	0.66	**1.88**	**0.63**
	90%	0.64	0.70	0.66	**2.23**	**0.75**
	100%	0.65	0.71	0.67	**2.25**	**0.76**
CLIP + K-Means (Online)	10%	**0.58**	**0.63**	**0.60**	**0.76**	**0.26**
	20%	0.59	**0.64**	**0.61**	**0.95**	**0.32**
	30%	0.59	**0.64**	**0.61**	**1.11**	**0.37**
	40%	0.60	**0.66**	**0.62**	**1.29**	**0.43**
	50%	0.61	0.67	0.63	**1.46**	**0.49**
	60%	0.63	0.69	0.65	**1.60**	**0.54**
	70%	0.64	0.70	0.66	**1.76**	**0.59**
	80%	0.64	0.70	0.66	**1.93**	**0.65**
	90%	0.64	0.71	0.66	**2.28**	**0.77**
	100%	0.65	0.71	0.67	**2.31**	**0.78**

Table 5.3: Performance comparison across label percentages for three models on **EXIST English test set** using **CLIP** Embeddings, including execution time and carbon emissions. **Bold and underlined** values indicate the best performance for each metric at a given label percentage. **Bold** values indicate results that are equal to or exceed the performance of the Finetuned CLIP baseline for the corresponding label percentage.

When switching from CLIP to **BLIP**, the overall behavior remains similar, but BLIP-based clustering shows even stronger relative gains at low supervision. For example, with 20–40% labeled data, both clustering variants achieve F1-scores around 0.56–0.59, closely matching

Model	% Labeled	P	R	F1	Time (min)	gCO_2
Finetuned BLIP	10%	0.57	0.59	0.56	2.85	2.16
	20%	0.57	0.60	0.58	5.09	3.86
	30%	0.61	0.66	0.63	7.12	5.39
	40%	0.61	0.64	0.61	9.62	7.29
	50%	0.63	0.66	0.63	11.50	8.71
	60%	0.62	0.66	0.63	14.90	11.30
	70%	0.62	0.66	0.63	15.74	11.90
	80%	0.61	0.65	0.62	18.98	14.40
	90%	0.62	0.66	0.63	19.83	15.00
	100%	0.63	0.67	0.64	22.31	16.10
BLIP + K-Means (Batch)	10%	0.54	**0.59**	**0.56**	**0.61**	**0.21**
	20%	0.54	0.59	0.56	**0.65**	**0.22**
	30%	0.55	0.60	0.57	**0.88**	**0.30**
	40%	0.55	0.60	0.57	**1.06**	**0.36**
	50%	0.55	0.60	0.57	**1.20**	**0.40**
	60%	0.55	0.60	0.57	**1.34**	**0.45**
	70%	0.56	0.61	0.58	**1.52**	**0.51**
	80%	0.56	0.62	0.59	**1.67**	**0.56**
	90%	0.56	0.62	0.59	**1.82**	**0.61**
	100%	0.57	0.62	0.59	**1.99**	**0.67**
BLIP + K-Means (Online)	10%	0.53	0.58	0.53	**0.64**	**0.22**
	20%	0.56	0.61	**0.58**	**0.79**	**0.27**
	30%	0.56	0.61	0.58	**0.92**	**0.31**
	40%	0.56	0.61	0.58	**1.10**	**0.37**
	50%	0.56	0.62	0.59	**1.24**	**0.42**
	60%	0.58	0.64	0.60	**1.39**	**0.47**
	70%	0.57	0.63	0.60	**1.57**	**0.53**
	80%	0.59	**0.65**	0.61	**1.72**	**0.58**
	90%	0.59	0.65	0.62	**1.88**	**0.63**
	100%	0.59	0.64	0.61	**2.05**	**0.69**

Table 5.4: Performance comparison across label percentages for three models on **EXIST English test set** using **BLIP** Embeddings, including execution time and carbon emissions. **Bold and underlined** values indicate the best performance for each metric at a given label percentage. **Bold** values indicate results that are equal to or exceed the performance of the Finetuned CLIP baseline for the corresponding label percentage.

or surpassing the finetuned BLIP model while requiring less than 1 gCO_2, a reduction of more than 90% compared to fine-tuning.

A direct comparison between CLIP- and BLIP-based models on the EXIST English dataset further clarifies the relative strengths of the two embedding families (see Figures 5.3 and

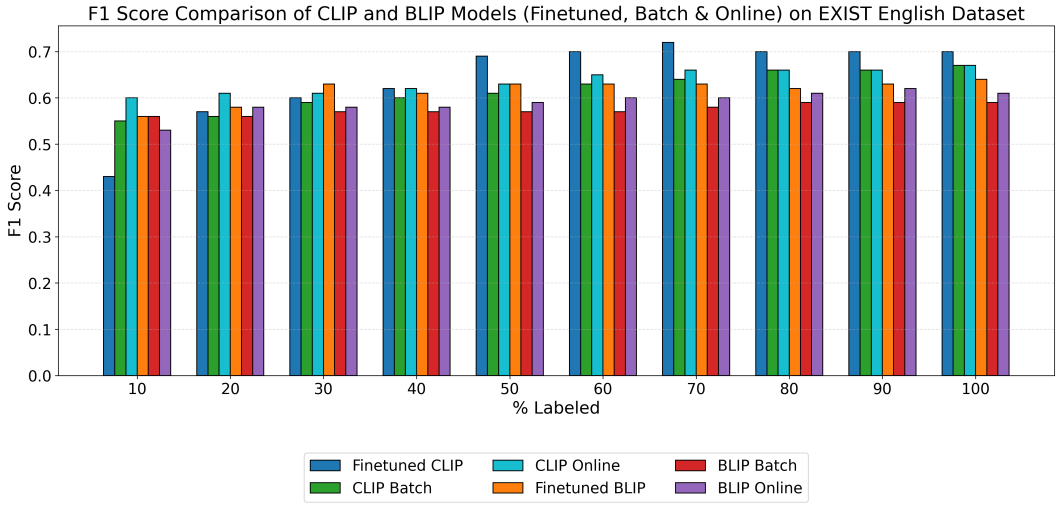


Figure 5.3: F1 scores obtained by finetuned, batch K-Means, and online K-Means variants of CLIP and BLIP embeddings across different proportions of labeled data on the EXIST English dataset. CLIP-based models consistently achieve higher performance than BLIP, while clustering-based approaches provide competitive results, particularly at lower label percentages.

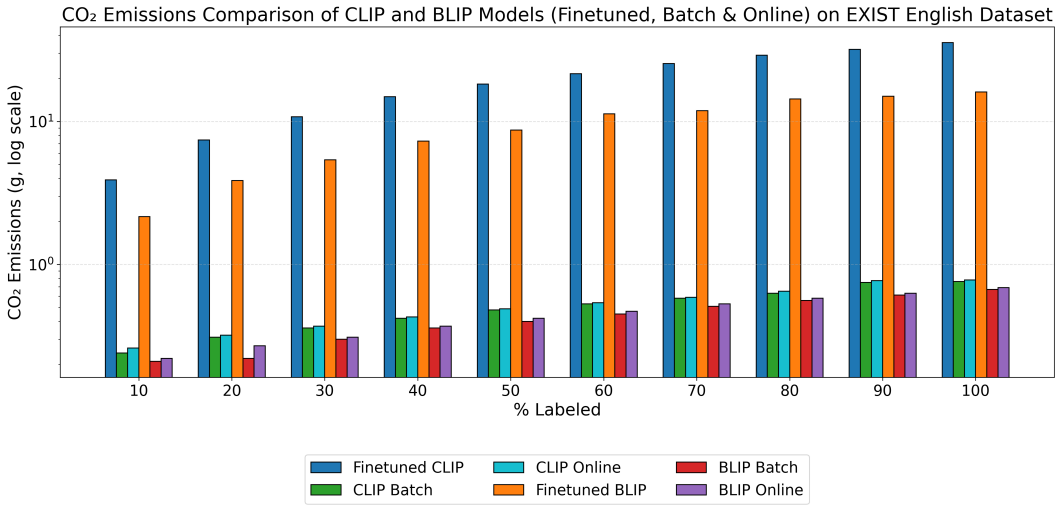


Figure 5.4: CO_2 emissions produced by finetuned, batch K-Means, and online K-Means variants of CLIP and BLIP across increasing label percentages on the EXIST English dataset. The logarithmic scale highlights the substantial environmental gap between finetuned models and the significantly more efficient clustering-based approaches.

5.4). Across all supervision levels, CLIP consistently achieves higher F1-scores than BLIP, both in the finetuned setting and under clustering. This gap is most visible at low supervision, where CLIP-based clustering reaches 0.60–0.62, while BLIP-based clustering remains around 0.56–0.59. However, BLIP compensates with slightly lower computational and environmental costs: finetuned BLIP emits roughly half the CO_2 of finetuned CLIP, and the BLIP clustering variants remain marginally more efficient than their CLIP counterparts. Overall, CLIP offers stronger predictive performance, whereas BLIP provides a more favorable efficiency profile. The clustering results highlight that this trade-off persists even without full supervision, reinforcing the value of evaluating both performance and sustainability jointly rather than in isolation.

Across the **EXIST Spanish experiments** (see Tables 5.5 and 5.6), the advantages of the clustering-based framework become even more pronounced. While the finetuned CLIP and BLIP models achieve moderate improvements with increasing supervision, their performance remains inconsistent, and the gains do not scale proportionally with the substantial rise in training time and emissions. In contrast, both batch and online K-Means variants deliver stronger and more stable F1-scores, reaching up to 0.62 for CLIP-based clustering and 0.59 for BLIP-based clustering. Notably, these results are achieved with emissions below 3 gCO_2, representing an order-of-magnitude reduction compared to the 14–36 gCO_2 required by finetuned models. This pattern indicates that, for Spanish, the clustering approaches not only preserve their efficiency advantage but also close the performance gap with full supervision more effectively than in English. The consistency of these trends across both embedding families reinforces the robustness of the proposed framework and highlights its suitability for multilingual abusive-language detection, particularly in scenarios where annotation resources or environmental constraints limit the feasibility of large-scale fine-tuning.

A closer comparison between CLIP and BLIP based models on the Spanish dataset (see Figures 5.5 and 5.6) further clarifies how the two embedding families behave under varying supervision. Similar to the English results, CLIP maintains a slight performance advantage: both its finetuned and clustering variants reach higher peak F1-scores, with CLIP-based clustering achieving 0.62 compared to BLIP's 0.59. However, the difference between the two models is smaller in Spanish, and BLIP-based clustering remains consistently competitive across all label percentages. In terms of efficiency, BLIP continues to offer a marginal environmental benefit, with both finetuned and clustering variants emitting slightly less CO_2 than their CLIP counterparts. Taken together, these trends show that while CLIP provides the strongest predictive performance, BLIP delivers a more favorable performance–efficiency balance in Spanish, reinforcing the importance of evaluating both accuracy and sustainability when selecting multilingual embedding models.

Model	% Labeled	P	R	F1	Time (min)	gCO_2
Finetuned CLIP	10%	0.49	0.54	0.48	5.1	3.86
	20%	0.54	0.56	0.50	9.76	7.39
	30%	0.53	0.58	0.54	14.5	10.9
	40%	0.55	0.58	0.53	19.02	14.4
	50%	0.50	0.55	0.50	23.84	18.1
	60%	0.59	0.59	0.53	28.38	21.5
	70%	0.59	0.60	0.54	33.1	25.1
	80%	0.60	0.61	0.56	38.25	29.0
	90%	0.59	0.63	0.59	42.52	32.2
	100%	0.61	0.61	0.56	46.94	35.6
CLIP + K-Means (Batch)	10%	0.55	0.63	**0.58**	**0.67**	**0.23**
	20%	0.55	0.63	**0.58**	**0.81**	**0.27**
	30%	0.56	0.64	**0.59**	**0.98**	**0.33**
	40%	0.57	0.64	**0.59**	**1.14**	**0.38**
	50%	0.57	0.65	**0.60**	**1.30**	**0.44**
	60%	0.57	0.65	**0.60**	**1.44**	**0.49**
	70%	0.57	0.65	**0.60**	**1.60**	**0.54**
	80%	0.57	0.64	**0.60**	**1.75**	**0.59**
	90%	0.58	0.66	**0.61**	**1.92**	**0.65**
	100%	0.59	0.67	**0.62**	**2.09**	**0.70**
CLIP + K-Means (Online)	10%	0.54	0.61	**0.57**	**0.70**	**0.24**
	20%	0.56	0.63	**0.59**	**0.85**	**0.29**
	30%	0.55	0.62	**0.58**	**1.03**	**0.35**
	40%	0.56	0.63	**0.59**	**1.18**	**0.40**
	50%	0.57	0.65	**0.60**	**1.34**	**0.45**
	60%	0.57	0.65	**0.60**	**1.49**	**0.50**
	70%	0.58	0.65	**0.61**	**1.65**	**0.56**
	80%	0.58	0.66	**0.61**	**1.81**	**0.61**
	90%	0.58	0.66	**0.61**	**1.98**	**0.67**
	100%	0.58	0.66	**0.61**	**2.15**	**0.72**

Table 5.5: Performance comparison across label percentages for three models on **EXIST Spanish test set** using **CLIP** Embeddings, including execution time and carbon emissions. **Bold and underlined** values indicate the best performance for each metric at a given label percentage. **Bold** values indicate results that are equal to or exceed the performance of the Finetuned CLIP baseline for the corresponding label percentage.

Across both datasets, our framework consistently improves with increased supervision and performs competitively even with limited labeled data. The online update variant shows particular strength in early supervision stages, suggesting its adaptability to incremental learning. These findings validate the generalizability of our method across diverse meme classification

Model	% Labeled	P	R	F1	Time (min)	gCO_2
Finetuned BLIP	10%	0.49	0.53	0.47	2.85	2.16
	20%	0.51	0.56	0.52	4.98	3.77
	30%	0.51	0.56	0.50	7.16	5.42
	40%	0.50	0.55	0.50	9.17	6.95
	50%	0.57	0.60	0.55	11.35	8.60
	60%	0.51	0.55	0.51	13.33	10.10
	70%	0.53	0.57	0.53	15.51	11.75
	80%	0.56	0.60	0.55	17.48	13.24
	90%	0.59	0.62	0.57	19.83	15.02
	100%	0.58	0.63	0.59	22.0	16.7
BLIP + K-Means (Batch)	10%	0.53	0.61	**0.56**	**0.72**	**0.24**
	20%	0.54	0.61	**0.57**	**0.83**	**0.28**
	30%	0.53	0.60	**0.56**	**0.99**	**0.33**
	40%	0.54	0.62	**0.57**	**1.14**	**0.38**
	50%	0.54	0.61	**0.57**	**1.24**	**0.42**
	60%	0.53	0.61	**0.56**	**1.38**	**0.46**
	70%	0.54	0.62	**0.58**	**1.53**	**0.52**
	80%	0.54	0.62	**0.58**	**1.70**	**0.57**
	90%	0.54	0.61	**0.57**	**1.85**	**0.62**
	100%	0.58	0.63	**0.59**	**2.21**	**0.74**
BLIP + K-Means (Online)	10%	0.52	0.59	**0.55**	**0.76**	**0.26**
	20%	0.53	0.60	**0.56**	**0.87**	**0.29**
	30%	0.53	0.60	**0.56**	**1.04**	**0.35**
	40%	0.53	0.61	**0.57**	**1.19**	**0.40**
	50%	0.54	0.61	**0.57**	**1.29**	**0.43**
	60%	0.54	0.61	**0.57**	**1.43**	**0.48**
	70%	0.54	0.62	**0.58**	**1.58**	**0.53**
	80%	0.55	0.63	**0.58**	**1.76**	**0.59**
	90%	0.55	0.62	**0.58**	**1.91**	**0.64**
	100%	0.55	0.62	**0.58**	**2.30**	**0.77**

Table 5.6: Performance comparison across label percentages for three models on **EXIST Spanish test set** using **BLIP** Embeddings, including execution time and carbon emissions. **Bold and underlined** values indicate the best performance for each metric at a given label percentage. **Bold** values indicate results that are equal to or exceed the performance of the Finetuned BLIP baseline for the corresponding label percentage.

tasks and languages, offering a practical alternative to fine-tuned models when annotation resources are constrained. They underscore the practicality of our approach: it avoids costly fine-tuning while still delivering competitive performance, making it ideal for low-resource or real-time deployment scenarios.

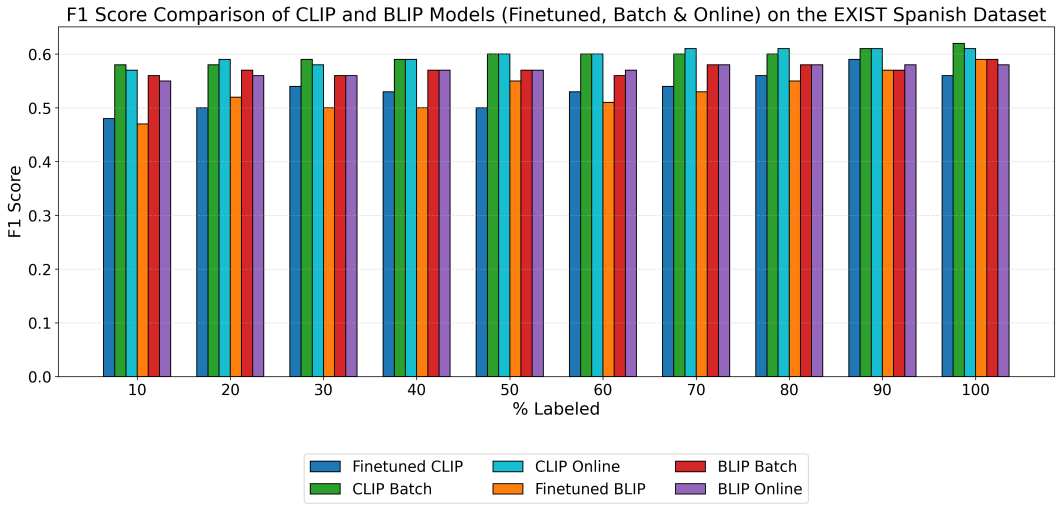


Figure 5.5: F1 scores obtained by finetuned, batch K-Means, and online K-Means variants of CLIP and BLIP embeddings across different proportions of labeled data on the EXIST Spanish dataset. Clustering-based approaches outperform the finetuned baselines at most label percentages, with both CLIP and BLIP clustering variants showing strong and stable performance across the labeling spectrum.

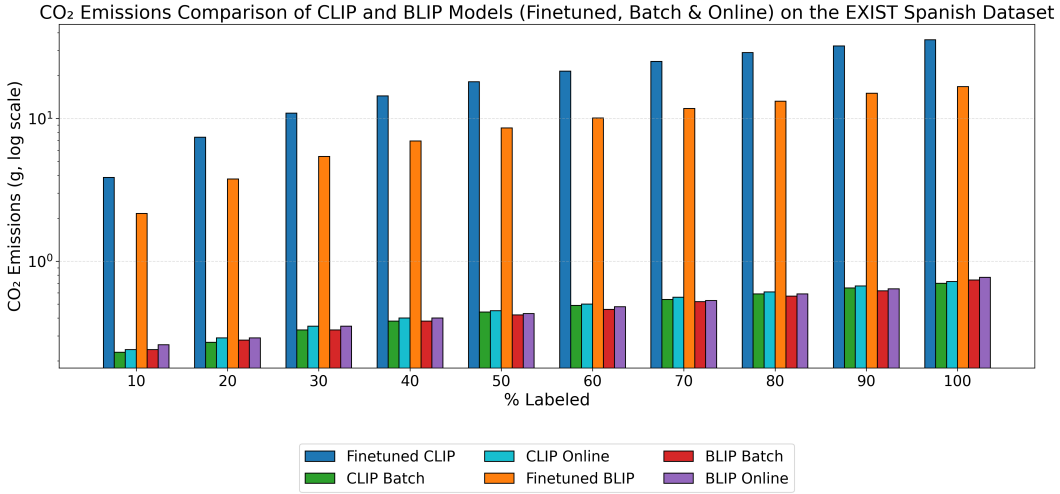


Figure 5.6: CO_2 emissions produced by finetuned, batch K-Means, and online K-Means variants of CLIP and BLIP across increasing label percentages on the EXIST Spanish dataset. The logarithmic scale highlights the substantial environmental advantage of clustering-based approaches, which maintain consistently low emissions compared to the significantly more carbon-intensive finetuned models.

5.4 Discussion

The findings underscore the dual advantage of our semi-supervised clustering framework: **data efficiency and sustainability**. Unlike finetuning, clustering does not require backpropagation or optimizer updates, and operates with small batch sizes. As a result, GPU utilization remains relatively low (40–50%), compared to finetuning (60—90%) . This difference translates directly into energy consumption: fine-tuned models require approximately 2–3× more GPU time per batch, leading to significantly higher carbon emissions. By quantifying emissions alongside accuracy, we provide a transparent framework for evaluating trade--offs between predictive performance and environmental impact.

Another important aspect is the robustness of our approach under limited supervision. At 30—40% labeled data, clustering achieves F1-scores comparable to fully fine-tuned baselines, while consuming 10–15× less carbon. This is a critical advantage in low-resource settings, where labeled data is scarce and sustainability is a priority. The ability to achieve strong results with less supervision and lower emissions positions our framework as a practical alternative to full finetuning.

We deliberately performed full finetuning rather than parameter-efficient fine-tuning (PEFT) methods such as LoRA [61] or QLoRA [38] to ensure a direct comparison between a fully fine-tuned supervised baseline and our semi-supervised approach. This avoids algorithmic differences that could confound performance comparisons.

Overall, the results highlight a clear trade–off between accuracy and environmental cost. While fine-tuned models achieve marginally higher performance at full supervision, our clustering framework delivers competitive accuracy with far less labeled data and dramatically lower emissions. This balance of efficiency, sustainability, and accuracy underscores the practical value of semi--supervised clustering for responsible AI development, and points toward future research directions that integrate PEFT with clustering to further reduce environmental impact.

While our semi-supervised clustering framework demonstrates clear advantages in terms of computational efficiency and environmental sustainability, it also presents several limitations. First, the performance of our approach remains slightly below that of fully supervised baselines, which benefit from extensive labeled data and fine-tuning. This performance gap can be attributed to the absence of gradient-based optimization on large annotated corpora and the reliance on pre-trained representations that may not fully capture the nuances of misogynistic or hateful multimodal content. Second, the constrained clustering formulation depends on the quality and representational alignment of the underlying vision–language

embeddings. In cases where the pre-trained model exhibits domain or cultural bias, these limitations may propagate to the clustering results.

The approach presented in this chapter treats annotation labels as ground truth, operating under the assumption that each instance has a single correct label. However, as discussed in Chapter 1, hateful content is inherently subjective i.e; human annotators often disagree about whether a given instance is misogynistic. This work does not model or account for such disagreement, focusing instead on establishing that efficient methods can achieve competitive hate detection performance. The question of whether annotation disagreement can be explicitly modeled within an equally efficient framework is addressed in the next work.

Chapter Summary

In this chapter, we discussed our proposed semi–supervised clustering framework with an embedding regularization strategy for misogyny and sexism detection in multimodal content. By leveraging pre–trained vision–language model embeddings and introducing constraint–based regularization, the method reduces the need for large–scale labeled datasets and fine–tuning, making it suitable for low–resource and multilingual environments. Despite achieving slightly lower classification performance compared to supervised baselines, our approach significantly lowers computational cost and carbon energy consumption. The analysis of energy–accuracy trade–offs reveals that modest reductions in accuracy can yield substantial environmental savings, reinforcing the value of lightweight and energy efficient models for socially responsible AI research. Overall, this study highlights a pathway toward greener multimodal learning, that balances performance with sustainability, paving the way for future systems that are both ethically and environmentally conscious.

6 Supervised Contrastive Learning Based Method to Model Hate and Disagreement

Building on the previous chapter's emphasis on efficiency and sustainability, this chapter shifts from a semi-supervised clustering paradigm to a supervised contrastive learning framework. While the clustering approach in Chapter 5 treated annotation labels as unambiguous ground truth, this chapter explicitly models the subjectivity inherent in hateful content by treating disagreement detection as a prediction task: asking whether annotators will agree or disagree on a given instance. The supervised framework offers greater expressive power through learned optimization while maintaining computational efficiency by operating on frozen pretrained embeddings, allowing us to examine whether both hate detection and disagreement detection can be achieved within the same lightweight paradigm.

The chapter proceeds in three stages: first, applying the contrastive framework to hate detection to establish its effectiveness on the core task; second, extending it to disagreement detection to demonstrate that content ambiguity can be predicted from the same pretrained representations; and third, combining both objectives in a joint prediction model that captures hate and disagreement simultaneously within a single efficient architecture.

Contributions

The primary objectives of this research are to (1) develop a multimodal classification framework for detecting misogynous memes, and (2) predict annotator disagreement within meme datasets, with a particular focus on identifying instances where annotators are likely to disagree on the label of misogynistic content. To achieve these goals, we contribute the following:

1. We introduce a pair-centric supervised contrastive objective that directly optimizes similarities between all label-consistent embedding pairs within a batch. The approach

facilitate the detection of misogynistic content as well as the identification of perceptual disagreement among users. This dual focus helps us better understand the varied perspectives found in online discussions, which are often missed by traditional models.

2. We design a contrastive learning framework in which we perform a systematic investigation of encoder architectures for both text and image modalities. This includes evaluating various combinations of pre-trained image and text encoders, as well as exploring different strategies for multimodal information fusion and representation aggregation. Our empirical analysis aims to identify the most effective configurations to improve performance on complex multimodal classification tasks.

3. We further employ a complementary projection-based architecture in which embeddings extracted from pre-trained vision–language models are refined using a lightweight Multilayer Perceptron (MLP) and optimized with the same contrastive objective. This formulation enables contrastive learning over unified multimodal embeddings while maintaining architectural simplicity and consistency across modeling paradigms.

4. We extend the proposed contrastive learning framework to the joint detection of hate and disagreement, enabling the model to learn coupled semantic attributes rather than a single label dimension. This formulation allows contrastive learning to operate over multi-attribute supervision and demonstrates the flexibility of our approach beyond binary misogyny classification.

6.1 Methodology

Our system employs a contrastive learning-based approach to predict misogyny and disagreement in memes. Contrastive learning is a framework that learns to distinguish between similar and dissimilar pairs by pulling embeddings of similar instances (similar class) closer together and pushing embeddings of dissimilar instances (different class) farther apart in the latent space (See section 2.4 for details).

Given a meme composed of visual and textual content, our framework first extracts feature representations using pretrained encoders. Depending on the representation strategy, multimodal embeddings are constructed either by explicitly combining unimodal image and text features or by leveraging pretrained vision–language models that produce joint multimodal representations. In both cases, the resulting embeddings are mapped to a shared latent space using a projection head and optimized using a contrastive learning objective. The proposed contrastive loss function encourages maximizing similarity between samples belonging to the same class while minimizing similarity across different classes.

After training, the learned embedding space is evaluated using a non-parametric K-Nearest Neighbors (KNN) classifier. For each test sample, embeddings are extracted using the same encoders and projection heads as in training. Cosine similarity is computed between each test embedding and all training embeddings, and the label of the test sample is determined by majority voting among its K most similar neighbors.

We use this architecture to address two key tasks in the domain of multimodal analysis:

1. **Misogyny Detection:** To classify memes into two categories: misogynistic and non-misogynistic. The goal was to identify memes containing language, imagery, or both that exhibit misogynistic intent.

2. **Disagreement Prediction:** This task focuses on identifying memes where annotators disagree on the label. Memes were classified into two categories:
   - **Agreement:** All annotators unanimously agreed on the label (either misogynistic or non-misogynistic)
   - **Disagreement:** Annotators provided conflicting labels, indicating ambiguity or subjectivity in meme interpretation

6.1.1 Fusion-Based Embedding Construction

This section describes the multimodal fusion strategy adopted to construct joint meme representations optimized using the contrastive learning objective introduced in Section 6.1.3. In this approach, images and text are first processed using dedicated encoders to extract embeddings for each modality (image and text). These embeddings are passed through a projection head to obtain the same dimensional embeddings for both modalities. After alignment, the image and text embeddings are combined using aggregation functions to create unified multimodal representations. The complete training architecture, including the encoding, aggregation, and optimization steps, is illustrated in Figure 6.1.

Text and Image Encoding

In this section, we outline the embedding strategies adopted for the multimodal meme data. Specifically, we describe how textual content is transformed into dense vector representations using state-of-the-art language models, followed by an explanation of how visual features are extracted from meme images using image encoders. These embeddings serve as foundational inputs for the downstream projection and alignment stages of our model.

Text Encoder In the proposed methodology, the textual part of memes is encoded using

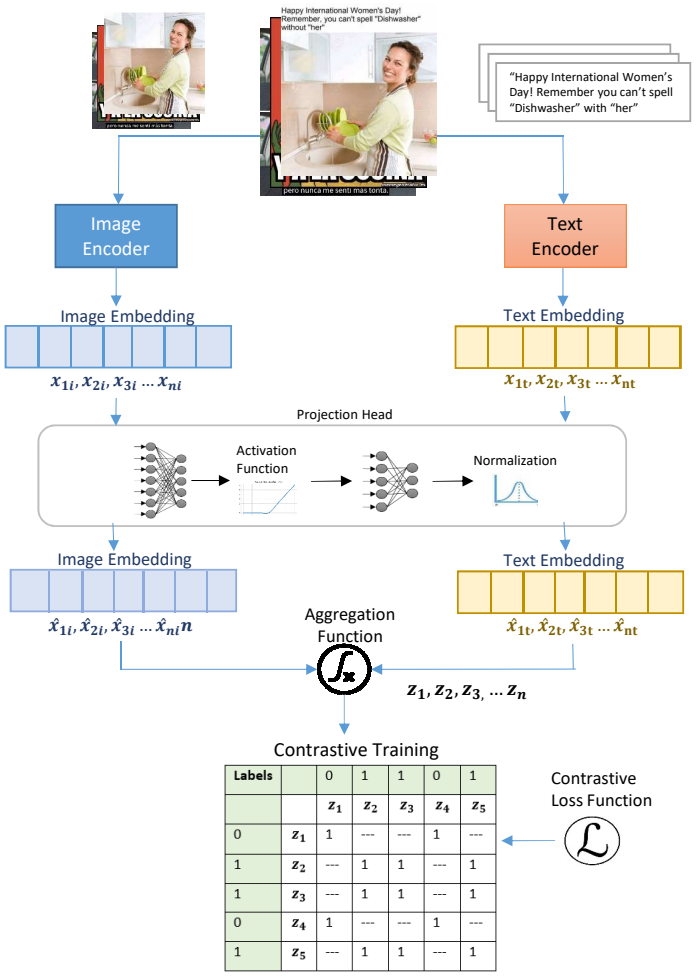


Figure 6.1: Overview of the fusion-based training architecture, illustrating feature extraction, embedding fusion, and contrastive optimization.

text encoders to get the text embeddings. We used the following two text encoders in our experiments:

- **BERT**: The first text encoder we used is a transformer-based architecture developed by [39]. It generates contextualized embeddings for each token in the input text.
- **RoBERTa**: We experimented with another text encoder named RoBERTa [89] to encode text into fixed-size feature vectors. RoBERTa is an optimized variant of BERT that is trained on a larger dataset. It uses dynamic masking during pretraining that results in more effective contextualized embeddings. This model processes each token through multiple transformer layers.

Given a meme m, the embedding representation of its text source will be denoted as x_{mt}.

Image Encoder The images from the memes dataset are also encoded using some pre-trained image models to convert the images to embeddings. We experimented using following models to get the image embeddings:

- **ResNet50**: The first model we used to get image embeddings is ResNet-50 developed by [57]. This deep convolutional neural network is renowned for its high performance in image recognition and other computer vision tasks. It includes 50 layers of convolution, batch normalization, and ReLU activation functions.

- **ViT**: We also experimented with Vision Transformer (ViT) [40] model to encode images into fixed-size feature vectors. The ViT processes input images by dividing them into patches and passing these patches through multiple transformer layers.

Analogously, for a meme m, the embedding representation of the image source will be denoted as x_{mi}.

Aggregation Functions

Once the embeddings generated for the text and image source have been derived, i.e., x_{mt} and x_{mi}, they are passed through a projection head in order to get the same dimensions and to transform the vectors into a lower dimensional space. To this purpose, we exploited a projection layer that maps both x_{mt} and x_{mi} to new representations, respectively $\hat{x}_{mt}$ and $\hat{x}_{mi}$, of the same dimensions. We therefore considered three different aggregation functions to join image and text embeddings.

Concatenation The text and image embeddings obtained from their respective encoders are aggregated to form a unified representation of each meme. To achieve this, we use simple concatenation, where the text embedding vector and the image embedding vector are joined end-to-end. This operation preserves the unique information from both modalities without altering their individual structures. The resulting concatenated vector captures both semantic and visual cues in a single, fixed-size representation.

Given a meme m, the concatenated feature vector z_m is formed by stacking the image and text embeddings along the feature axis as:

$$z_m = \hat{x}_{mi} || \hat{x}_{mt} \tag{6.1}$$

z_m has dimensions $(n + k)$, where n is the dimensionality of the image embeddings, and k is the dimensionality of the text embeddings.

Hadamard Product To aggregate the text and image embeddings into a unified representation, we also adopted the Hadamard product. This operation combines the two vectors by multiplying their corresponding elements, emphasizing the features that are mutually strong in both modalities. Unlike concatenation, the Hadamard product results in a vector of the same dimensionality as the original embeddings. It encourages interaction between aligned dimensions of the text and image representations. The resulting feature vector z is computed as:

$$z_m = \hat{x}_{mi} \odot \hat{x}_{mt} \tag{6.2}$$

The resulting feature vector z_m is an n-dimensional vector, where each coordinate is a combination of the respective image and text embedding features through their element-wise product.

Consistency and Complementarity To effectively integrate multimodal information, we adapted the feature fusion strategy proposed by [45], which aims to capture both consistency and complementarity between embeddings.

Consistency, implemented using the Hadamard product, again emphasizes the shared information between the image and text embeddings. By multiplying corresponding elements of the two vectors, this operator highlights dimensions where both modalities exhibit strong, mutually reinforcing features. Specifically, the consistency component z_m^{con} is derived by applying the Hadamard product between the projected image and text embeddings:

$$z_m^{con} = \hat{x}_{mi} \odot \hat{x}_{mt} \tag{6.3}$$

Complementary, implemented via element-wise sum, captures the distinct and supplementary information present in the image and text embeddings. While each modality may encode unique features, such as visual style in images or semantic nuance in text, complementarity preserves and merges these non-overlapping aspects. By summing the vectors, the model emphasizes diverse features that are not necessarily aligned across modalities but still contribute meaningfully to the overall representation. The complementary component results in a feature vector z_m^{com}:

$$z_m^{com} = \hat{x}_{mi} \oplus \hat{x}_{mt} \tag{6.4}$$

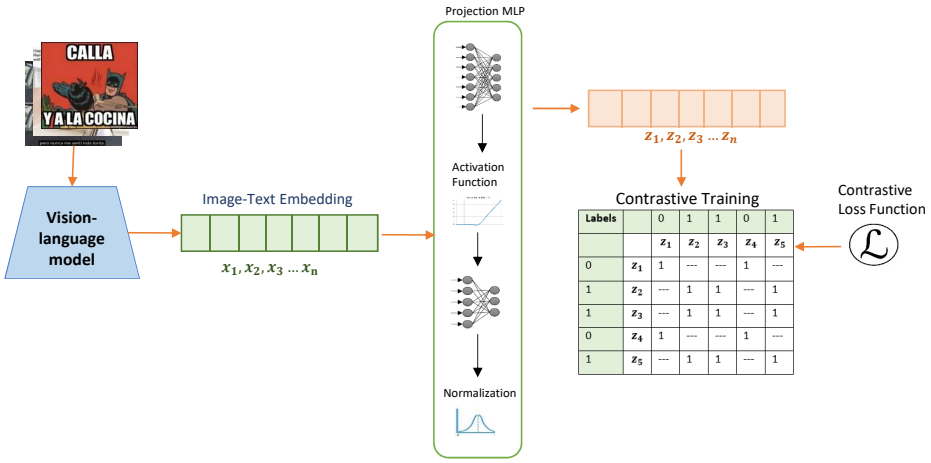


Figure 6.2: Overview of the MLP-based training architecture.

Finally, the embedding vector z determined according to the consistency and complementary representations are aggregated by summing their contribution:

$$z_m = z_m^{con} \oplus z_m^{com} \tag{6.5}$$

6.1.2 MLP-Based Embedding Projection

This section describes the MLP-based representation learning strategy used to construct multimodal meme embeddings from pretrained vision–language models. Unlike the fusion-based approach, which explicitly aggregates unimodal image and text features, this method leverages pretrained vision-language models to obtain joint vision–language representations. These pretrained embeddings are subsequently refined using a lightweight projection network and optimized using the contrastive learning objective described in Section 6.1.3. The complete architecture is shown in Figure 6.2.

Pretrained Multimodal Feature Extraction

Each meme in the datasets contains both visual and textual elements. To represent these modalities, we experimented with two vision–language models: CLIP [117] and BLIP [81]. For CLIP, the processor tokenizes the text and normalizes the images to ensure alignment, generating separate embeddings for each modality. These embeddings are then combined to produce a single multimodal feature vector. In parallel, BLIP is employed to extract embeddings that model fine-grained interactions between vision and language. we use the embeddings from each model independently to train our clustering algorithms.

Projection Network

The MLP-based architecture employs a lightweight projection network designed to transform pretrained CLIP/BLIP embeddings into a compact representation suitable for contrastive learning. The network follows a simple feed-forward structure consisting of two fully connected layers with a non-linear activation in between.

To ensure stability during contrastive optimization, the output of the network is L2-normalized, producing embeddings that lie on the unit hypersphere. In short, given an input embedding x_m from pretrained CLIP/BLIP model, the MLP produces a normalized representation:

$$z_m = \mathrm{MLP}(x_m) \tag{6.6}$$

which is then trained using the contrastive learning objective introduced in Section 3.3.

6.1.3 Contrastive Learning

The goal of contrastive learning is to increase the similarity between similar samples and decrease the similarity between dissimilar ones. The original **Supervised Contrastive (SupCon) Loss**, introduced by [71], extends the self-supervised contrastive learning framework to the supervised setting. Using cosine similarity to measure relationships between embeddings, the SupCon loss function is defined as:

$$\mathcal{L}_{\text{out}}^{\text{sup}} = \sum_{m \in M} L_{\text{out},m}^{\text{sup}} = \sum_{m \in M} \frac{-1}{|M_m^+|} \sum_{p \in M_m^+} \log \frac{\exp(z_m \cdot z_p / \tau)}{\sum\limits_{a \in M : a \neq m} \exp(z_m \cdot z_a / \tau)} \tag{6.7}$$

where $z_m \cdot z_p$ corresponds to the cosine between them normalized by their norm product, M represents the set of all anchors in the batch, M_m^+ is the set of all positive samples in the batch for each anchor m, and τ is a temperature parameter.

In meme understanding tasks, samples belonging to the same class often exhibit high intra-class diversity across both textual and visual modalities. In addition, class imbalance is common, particularly for subjective categories such as misogyny and disagreement. Under these conditions, anchor-based contrastive objectives may bias optimization toward dominant or easier anchors, as the loss is computed independently for each anchor and normalized at the sample level. This can limit the ability of the model to capture class-level semantic cohesion across heterogeneous multimodal samples. In this paper, we propose a variation of the original loss function, which is based on a **no-anchor** approach. Our formulation is inspired by

recent efforts toward contrastive objectives [62], [12] and [21], but differs fundamentally in both formulation and application domain.

Let $\mathcal{P}$ be the set of all **unique positive pairs** in the batch with class label y, i.e, $\mathcal{P} = \{\langle z_m, z_k\rangle \mid k < m \text{ and } y_m = y_k\}$. A pair $\langle z_m, z_k\rangle$ is considered positive if the samples z_m and z_k share the same label (i.e., $y_m = y_k$). Similarly, let $\mathcal{A}$ denote the set of **all unique pairs** in the batch, i.e., $\mathcal{A} = \{\langle z_r, z_s\rangle \mid r < s\}$. The **Pairwise Global Supervised Contrastive (PaG-SCon)** Loss function $L_{\text{pg}}^{\text{sup}}$ is then defined as:

$$\mathcal{L}_{\text{pg}}^{\text{sup}} = -\frac{1}{|\mathcal{P}|} \sum_{<z_m, z_k> \in \mathcal{P}} \log \frac{\exp(z_m \cdot z_k)/\tau)}{\sum\limits_{<z_r, z_s> \in \mathcal{A}} \exp(z_r \cdot z_s)/\tau)} \tag{6.8}$$

The denominator acts as a normalization factor and is defined as the sum of exponential cosine similarities for all unique pairs in the batch. Unlike the original SupCon loss, which computes the loss for each anchor individually by considering its positive pairs and all pairs, PaG-SCon eliminates the need for anchors altogether. Instead, we consider all positive pairs in the batch and normalize the loss by the total number of unique positive pairs $\mathcal{P}$. This formulation simplifies the computation and ensures that the loss is computed globally across the batch, rather than being anchored to individual samples.

The gradient with respect to z_m is:

$$\frac{\partial \mathcal{L}_{\text{pg}}^{\text{sup}}}{\partial z_m} = -\frac{1}{\tau} z_k + \frac{1}{\tau} \cdot \frac{\sum\limits_{\langle z_r, z_s\rangle \in \mathcal{A}} \exp(z_r \cdot z_s/\tau) \cdot \frac{\partial (z_r \cdot z_s)}{\partial z_m}}{\sum\limits_{\langle z_r, z_s\rangle \in \mathcal{A}} \exp(z_r \cdot z_s/\tau)} \tag{6.9}$$

In our case, only those pairs $\langle z_r, z_s\rangle$ where either $z_r = z_m$ or $z_s = z_m$ will contribute non-zero gradients. Therefore, we can rewrite Equation 6.9 as follows:

$$\frac{\partial \mathcal{L}_{\text{pg}}^{\text{sup}}}{\partial z_m} = \frac{1}{|\mathcal{P}|} \sum_{\langle z_m, z_k\rangle \in \mathcal{P}} \frac{1}{\tau} \left[-z_k + \frac{1}{D} \left(\sum_{\substack{\langle z_r, z_s\rangle \in \mathcal{A} \\ z_r = z_m}} \exp(z_m \cdot z_s/\tau) z_s + \sum_{\substack{\langle z_r, z_s\rangle \in \mathcal{A} \\ z_s = z_m}} \exp(z_r \cdot z_m/\tau) z_r \right) \right] \tag{6.10}$$

where

$$D = \sum_{\langle z_r, z_s\rangle \in \mathcal{A}} \exp(z_r \cdot z_s/\tau) \tag{6.11}$$

The term $-\frac{1}{\tau} z_k$ acts as a negative attractive force to align the two vectors by minimizing the angle between z_m and z_k, i.e., increasing their similarity. The term $\frac{1}{\tau D} \sum\limits_{\substack{\langle z_r, z_s\rangle \in \mathcal{A} \\ z_r = z_m}} \exp(z_m \cdot$

$z_s/\tau)z_s + \sum_{\substack{\langle z_r, z_s\rangle \in \mathcal{A} \\ z_s = z_m}} \exp(z_r \cdot z_m/\tau)z_r$ represents a positive repulsive force from all pairs involving $z_m \in \mathcal{A}$ plays the role of positive repulsive force that pushes the embedding z_m away from everything else.

The proposed function differs from that of the original SupCon. In our formulation, the softmax probabilities are normalized over all pairs in the batch, while in SupCon they are normalized over all samples in the batch with respect to each anchor z_m. The proposed loss, and therefore its gradient, uses a uniform weight $\frac{1}{|\mathcal{P}|}$ for all positive pairs, while SupCon uses a uniform weight but specific to the positives of z_m.

6.2 Experimental Settings

6.2.1 Datasets

We evaluate our proposed method using two publicly available benchmark datasets EXIST and MAMI for sexism and misogyny detection in memes (see Section 4.1). For each dataset, we used the official test sets provided to assess the generalization performance of our clustering framework. For Misogyny/Sexism detection task, the original test set was reserved for evaluation, while the training was performed on original training data. In contrast, the Disagreement detection task was evaluated using 10-fold cross-validation. For EXIST dataset, both tasks are evaluated using 10 fold cross validation

6.2.2 Baselines

To comprehensively evaluate our approach, we compare it against two categories of baseline models. The first category consists of fine-tuned multimodal baselines, where pre-trained vision–language models are used for the misogyny/sexism and disagreement detection tasks under identical training conditions. These baselines assess how well strong pretrained multimodal encoders perform when finetuned on our datasets. The second category features a state-of-the-art method from existing literature, offering a strong benchmark and placing our results in the wider context of multimodal meme analysis.

Fine-Tuned Multimodal Baselines

We implemented two widely recognized multimodal models, CLIP [117] and BLIP [81], as baselines on the same datasets used in our approach. Both baselines were finetuned and tested under identical conditions (dataset splits, preprocessing pipelines, evaluation protocols) to allow direct and fair comparison.

CLIP: CLIP is a state-of-the-art multimodal framework that jointly learns representations of text and images (See Section 4.2.1 for details). For this task, each meme, consisting of an image and its associated text, was processed using the CLIP processor to tokenize the text and preprocess the image. The processor ensures alignment between the visual and textual features and standardizes the input format through truncation and padding. A classification head is added to the model that consists of a fully connected layer mapping the combined image and text embeddings to binary logits, corresponding to the two target classes. The image and text embeddings produced by CLIP were combined to create a unified multimodal representation before being passed to the classifier.

BLIP: BLIP is a state-of-the-art framework for multimodal tasks that effectively combines visual and textual information (See Section 4.2.2 for details). The model was fine-tuned on respective datasets for misogyny/sexism and disagreement detection tasks. For preprocessing, the BLIP processor was used to combine visual and text features from the memes. It ensures consistency in the length and format of the text through truncation and padding. For fine-tuning, we adapted the BLIP model to our specific task by including a classification head, allowing it to predict binary labels. To optimize training efficiency, only the last few layers of the BLIP model were unfrozen, enabling partial fine-tuning while leveraging pretrained weights.

State-of-the-Art Comparison Model

We consider recent representative approaches for harmful and misogynous meme detection that report results on the MAMI dataset. [63] explores the use of large multimodal language model agents and achieves strong macro-F1 performance; however, its reliance on external proprietary models and agent-based inference introduces methodological differences from end-to-end trainable frameworks such as ours. Knowledge-enhanced approaches such as KERMIT [55] extend the task formulation by incorporating external knowledge or rationales, making them conceptually distinct from representation-learning-focused methods. A latest study on misogyny detection using MAMI dataset by [93] provided comparable performance but due to the unavailability of reproducible implementation at the time of our experiments, we did not perform experiments. Finally, we include the work by [125] as a direct baseline due to its alignment with the task definition and evaluation protocol used in this study.

Rizzi et al. [125]: As a state-of-the-art baseline, we include the Multimodal Text and Tags (MTT) model proposed by [125]. All experiments were reproduced under identical conditions and on the same datasets for both misogyny and disagreement detection tasks. The

original work provides a comprehensive evaluation of unimodal and multimodal architectures for automated misogyny detection in memes. In particular, early fusion architectures for multimodal approaches, such as Multimodal Text and Tags (MTT) and Multimodal Text and Caption (MTC), demonstrate superior performance by concatenating complementary features. The MTT model used in our evaluation employs a joint multimodal representation construction, integrating the meme text embedding with a visual tag embedding. Text transcriptions are encoded into a dense vector using a pretrained language model, while visual information is abstracted through tags representing salient objects or attributes in the image. In particular, visual tags are extracted using Clarifai API, a pretrained image recognition service, and exploited in the definition of a dense feature vector. The concatenated text and tag embeddings are fed into a feed-forward neural network comprising a fully connected layer (1024 neurons), a hidden layer (512 neurons with LeakyReLU and dropout), and a sigmoid output for binary classification.

6.2.3 Fusion-Based Embedding Generation and Aggregation

For Fusion based method, we employed two pre-trained text encoders, BERT [39] and RoBERTa [89], to extract embeddings from textual data. Each text sample was tokenized using the respective tokenizer for the model being used, ensuring that sequences were padded and truncated to the specified maximum length. The tokenized text data, along with attention masks and labels, were converted into tensors suitable for input to the respective models. The models were initialized with pre-trained weights, and embeddings were extracted by processing the tokenized inputs through the respective architectures. For both models, the hidden state corresponding to the special [CLS] token from the final layer was used as the sentence-level representation. This vector, a 768-dimensional embedding, captures the semantic meaning of the input text and serves as the basis for downstream tasks.

For image encoding, we utilized ResNet-50 [57] and Vision Transformer (ViT) [40], to extract fixed-size feature vectors from the input meme images. For ResNet-50, the input images were resized to 224 × 224 and preprocessed to meet the respective input requirements. We initialized a pre-trained ResNet-50 model from the torchvision library [99]. The classification layer of ResNet-50 was replaced with an Identity layer, allowing us to extract the 2048-dimensional feature vector. The entire model was made trainable to maximize adaptability to our dataset. For ViT, we used the ViTForImageClassification model from the Hugging Face transformers library [152]. The output corresponding to the [CLS] token from the last hidden state was extracted as the image embedding, resulting in a 768-dimensional feature vector. All parameters of the ViT model were initialized with pre-trained weights, and

their trainability was controlled based on the experimental setup. In our implementation, fine-tuning was enabled for all ViT layers during training.

Since the original image and text embeddings have different dimensions, they were first passed through a projection layer to reduce their dimensions to a unified 256-dimensional space. This projection layer consists of a Linear projection layer, Gaussian Error Linear Unit (GELU) activation function [58] and fully connected layers. It ensures that the image and text embeddings are aligned to the same dimension to facilitate the downstream tasks.

The resulting 256-dimensional image and text embeddings were combined using various fusion strategies, including concatenation, hadamard product and consistency and complementarity [45] aggregation methods. In our proposed approach, we explored multiple configurations for fusing the embeddings from text and image encoders. The following combinations of text and image encoders were utilized:

1. BERT with ResNet50
2. BERT with ViT
3. RoBERTa with ResNet50
4. RoBERTa with ViT

These fused embeddings, which maintain a 256 × 256-dimensional structure, were then used for downstream classification tasks i.e., predicting misogyny and disagreement in memes.

6.2.4 MLP-Based Embedding Generation

In the experimental setup, embeddings extracted from the pretrained CLIP or BLIP models are passed through a lightweight MLP-based projection module. The module takes as input a fixed-dimensional representation produced by the vision–language encoder and transforms it into a compact latent space suitable for contrastive learning. Specifically, the projection network consists of an input linear layer followed by a hidden layer with 512 units and a ReLU activation, enabling non-linear transformation of the pretrained embeddings. The resulting hidden representation is then mapped through a final linear layer to a lower-dimensional projection space with dimensionality set to 256.

The output of the projection network is L2-normalized to ensure that all embeddings lie on the unit hypersphere, which stabilizes contrastive optimization and allows similarity to be computed using cosine distance. This projected and normalized embedding is subsequently used in all contrastive learning experiments.

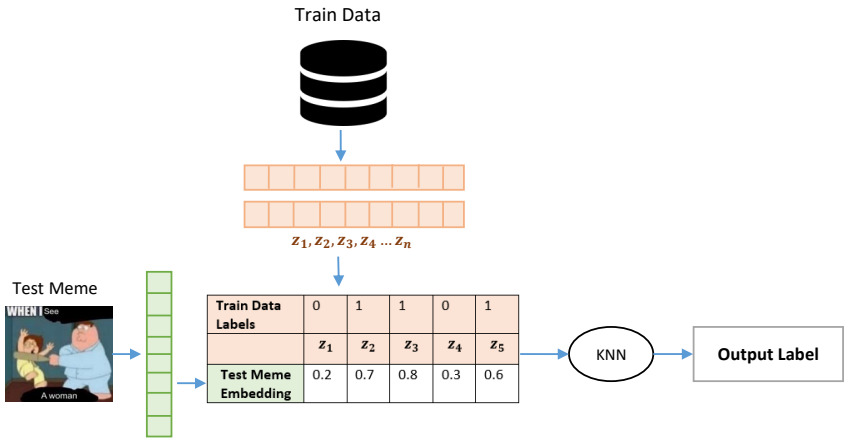


Figure 6.3: KNN-based evaluation procedure for the proposed contrastive learning framework.

6.2.5 Training and Inference

Our implementation was developed using the PyTorch [110] and Transformers [152] libraries in Python. For training with fusion based method, the aggregated image text embeddings are trained using contrastive loss function with the learning rates of 10^{-5} for both image and text encoders. Adam optimizer was employed with a batch size of 32. A dropout layer was added to mitigate overfitting and enhance regularization.

For the MLP-based projection network, the network is trained for 40 epochs using the Adam optimizer with a learning rate of 10^{-3}. Training is performed with batch size of 64, and gradients are updated after each batch.

After training, the model is tested by calculating the cosine similarity between each test sample and all training samples. Using the K-Nearest Neighbors (KNN) [31], 11 training embeddings with the highest cosine similarity to each test sample were extracted. The label for each test sample was determined by selecting the most frequent label among these 11 training samples. The testing procedure is illustrated in Figure 6.3.

6.3 Results and Discussion

This section presents a detailed analysis of the experimental results obtained for misogyny/sexism and disagreement detection in multimodal memes. We evaluate the effectiveness of the proposed contrastive learning framework across different embedding construction strategies, including explicit multimodal fusion and contrastive MLP-based representations learned from pretrained vision–language models. The experiments are conducted on MAMI and EXIST using standard evaluation metrics: Precision (P), Recall (R), and F1-score (F1) for both

Method	Aggregation	Models	P-	P+	R-	R+	F1-	F1+	F1-macro
C-Fusion	Concatenation	BERT+ResNet50	0.71	0.65	0.59	0.76	0.65	0.70	0.67
		BERT+ViT	0.83	0.61	0.43	0.91	0.56	0.73	0.65
		RoBERTa+ResNet50	0.66	0.67	**0.69**	0.64	0.67	0.66	0.66
		RoBERTa+ViT	0.85	0.60	0.36	0.94	0.51	0.73	0.62
	Hadamard Product	BERT+ResNet50	0.74	0.64	0.55	0.80	0.63	0.71	0.67
		BERT+ViT	0.81	**0.68**	0.59	0.87	**0.68**	**0.76**	**0.72**
		RoBERTa+ResNet50	0.71	0.66	0.60	0.76	0.65	0.70	0.68
		RoBERTa+ViT	0.77	0.60	0.42	0.88	0.54	0.71	0.63
	Consistency & Complementarity	BERT+ResNet50	0.73	0.67	0.62	0.77	0.67	0.72	0.69
		BERT+ViT	0.84	0.63	0.48	0.91	0.61	0.75	0.68
		RoBERTa+ResNet50	0.69	0.65	0.61	0.73	0.65	0.69	0.67
		RoBERTa+ViT	**0.87**	0.62	0.44	**0.93**	0.58	0.75	0.66
C-MLP	—	CLIP	**0.88**	**0.70**	**0.61**	0.92	**0.72**	**0.80**	**0.76**
	—	BLIP	0.80	0.61	0.44	0.89	0.57	0.73	0.65
Finetuning	—	CLIP	0.87	0.65	0.51	0.92	0.64	0.76	0.70
	—	BLIP	0.85	0.61	0.42	0.90	0.56	0.74	0.65
Rizzi et al. [125]	—	MTT	0.80	0.67	0.57	0.86	0.67	0.75	0.71

Table 6.1: Model performance on **Misogyny Detection** task on **MAMI** Dataset using 10,000 memes for training and 1,000 memes for testing on Test Data. Results are grouped by method: C-Fusion, C-MLP, and supervised baselines (Finetuning, Rizzi et al. [125]). Within each Method block, **bold** values indicate the best results that outperform all corresponding baseline models. **bold and underlined** values denote the best *F1-macro* score achieved within that method.

classes: hate/disagreement (+) and non-hate/agreement (–). Macro-averaged F1 is used as the primary measure to ensure balanced performance across classes.

6.3.1 Misogyny/Sexism Detection

To evaluate the proposed method, we conducted experiments on misogyny/sexism detection using the MAMI and EXIST datasets. For MAMI dataset, the model was trained on the training set comprising 10,000 memes and evaluated on a test set of 1,000 memes. For the EXIST dataset, a 10-fold cross-validation approach was employed to assess performance.

On the **MAMI dataset** (see Table 6.1), fusion-based contrastive learning demonstrates that aggregation strategy plays a significant role in overall effectiveness. The concatenation aggregation function yields moderate performance, it typically produces a high recall for the misogynous class and a lower recall for the non-misogynous class, especially when ViT-based visual features are utilized. Hadamard product aggregation consistently improves over concatenation by enforcing stronger feature interactions between modalities. With a macro-F1 score of

Method	Aggregation	Models	P-	P+	R-	R+	F1-	F1+	F1-macro
C-Fusion	Concatenation	BERT+ResNet50	0.68	0.70	0.55	**0.79**	0.61	**0.74**	**<u>0.68</u>***†
		BERT+ViT	0.63	0.69	**0.57**	0.75	0.60	0.72	0.66†
		RoBERTa+ResNet50	0.62	0.67	0.52	0.75	0.56	0.71	0.64
		RoBERTa+ViT	0.58	0.65	0.51	0.71	0.54	0.68	0.61
	Hadamard Product	BERT+ResNet50	0.52	0.62	0.45	0.67	0.49	0.64	0.57
		BERT+ViT	0.60	0.67	0.49	0.75	0.53	0.70	0.62
		RoBERTa+ResNet50	0.52	0.62	0.46	0.68	0.48	0.65	0.57
		RoBERTa+ViT	0.53	0.60	0.38	0.73	0.43	0.66	0.55
	Consistency & Complementarity	BERT+ResNet50	0.65	0.69	0.53	0.77	0.58	0.73	0.66*
		BERT+ViT	0.68	0.70	0.56	**0.79**	0.61	0.74	**<u>0.68</u>***†
		RoBERTa+ResNet50	0.53	0.62	0.46	0.68	0.49	0.65	0.57
		RoBERTa+ViT	0.56	0.63	0.46	0.71	0.50	0.67	0.59
C-MLP	—	CLIP	0.61	**0.75**	0.48	0.84	0.54	**0.79**	**<u>0.67</u>***†
	—	BLIP	0.54	0.72	0.43	0.80	0.48	0.76	0.62
Finetuning	—	CLIP	0.58	0.69	0.51	0.74	0.52	0.71	0.62
	—	BLIP	0.58	0.70	0.56	0.72	0.57	0.71	0.64
Rizzi et al. [125]	—	MTT	0.67	0.72	0.52	0.82	0.59	0.77	0.68

Table 6.2: Performance on the **Sexism Detection** task of the **EXIST** dataset using 10-fold cross-validation on the training set. Methods are grouped into *C-Fusion*, *C-MLP*, and supervised baselines (Finetuning, Rizzi et al. [125]). Within each Method block, **bold** values indicate the best results that outperform all corresponding baseline models, while **<u>bold and underlined</u>** values denote the best *F1-macro* score within that method. Symbols (*) and (†) indicate statistically significant improvements over Finetuned CLIP and BLIP, respectively.

0.72, the BERT + ViT setup with Hadamard product specifically gets the best fusion-based performance. Aggregation strategies based on consistency and complementarity further enhance robustness by explicitly modeling misogyny and non-misogyny between modalities, although their gains remain incremental compared to the best-performing fusion setup.

Despite these improvements, fusion-based methods are limited by their dependence on manually created aggregation functions and independently pretrained unimodal encoders. This limitation becomes evident when compared to the MLP-based contrastive learning approach. Leveraging pretrained vision–language models, the contrastive MLP trained on CLIP embeddings achieves the best overall performance on MAMI, with a macro-F1 score of 0.76. Importantly, this model exhibits balanced performance across classes, achieving high F1 scores for both misogynous and non-misogynous memes. It should also be noted that the obtained results are comparable with the F1-score previously reported for the MAMI dataset in [93], further supporting the robustness of the proposed approach. In contrast, the BLIP-based

(a) Meme from MAMI Dataset having True label "misogynous" misclassified by all models

(b) Meme from MAMI Dataset having True label "non-misogynous" but misclassified by all models

(c) Meme with label "non-misogynous" correctly classified only by BERT+ViT (hadamard product)

Figure 6.4: Illustration of Memes on Misogyny/Sexism Detection

MLP model yields lower gains, suggesting that the quality of multimodal alignment learned during pretraining directly impacts the effectiveness of contrastive fine-tuning.

A similar trend is observed on the **EXIST dataset** (see Table 6.2), evaluated using 10-fold cross-validation. Fusion-based contrastive models again show competitive performance, with the best results achieved using BERT-based textual encoders combined with ResNet50 or ViT visual features. In particular, concatenation and consistency & complementarity aggregation strategies achieve macro-F1 scores up to 0.68, matching the reported state-of-the-art MTT model. These configurations also show statistically significant improvements over finetuned CLIP and BLIP baselines, indicating that contrastive supervision contributes positively even when using explicit fusion.

However, unlike MAMI, the performance gap between fusion-based and MLP-based methods on EXIST is narrower. The contrastive MLP using CLIP achieves a macro-F1 score of 0.67, comparable to the strongest fusion configurations and the MTT model. While this model excels in identifying sexist content, as reflected by high recall and F1-score for the positive class, its performance on the non-sexist class remains comparatively lower. This suggests that the linguistic diversity and subtler forms of sexism present in EXIST pose additional challenges, particularly for models that rely heavily on global multimodal representations.

Across both datasets, direct fine-tuning of vision–language models consistently underperforms contrastive learning-based approaches. This trend highlights the benefit of contrastive objectives in structuring the embedding space by explicitly enforcing inter-sample similarity and dissimilarity constraints. The findings also demonstrate that learning representations directly in a shared multimodal space offers better generalization and more balanced predictions.

Figure 6.4 illustrates the examples of memes from MAMI and EXIST datasets. In Figure 6.4(a), the meme features a man alongside non-misogynous text. However, when the text is paired with the image, it shows a misogynous meme. This meme is misclassified as non-misogynous by all models, including baselines and the proposed models. This example demonstrates the challenge of interpreting the interplay between text and image. Figure 6.4(b) shows a woman in the image. The true label of this meme is non-misogynous, but all the models, including the baselines and proposed methods, misclassify it as misogynous. This highlights the difficulty models face in handling subtle contextual cues that distinguish misogynous from non-misogynous memes. The meme in Figure 6.4(c) is correctly classified only by the proposed best-performing model, which uses BERT and ViT encoders combined with the Hadamard product.

6.3.2 Disagreement Detection

To investigate annotator disagreement in multimodal memes, experiments were conducted on the MAMI and EXIST datasets following a 10-fold cross-validation strategy. Strong class imbalance and the subtle semantic cues that differentiate agreement from disagreement in multimodal memes make the disagreement detection more difficult than the misogyny/sexism detection task. The results demonstrate that contrastive learning consistently outperforms traditional fine-tuning across both the MAMI and EXIST datasets, while the absolute performance gains are relatively smaller, which relects the inherent difficulty of the task.

On the **MAMI dataset** (see Table 6.3), fusion-based contrastive models achieve moderate macro-F1 scores in the range of 0.51–0.55. Concatenation-based fusion generally performs competitively, with the BERT + ResNet50 configuration achieving the strongest fusion result (macro-F1 = 0.55), which is statistically significant over both finetuned CLIP and BLIP baselines. These models show a significant bias toward majority-class prediction, with high recall and F1 scores for the agreement class but significantly lower recall for the disagreement class.

Hadamard product aggregation slightly improves the modeling of minority-class samples in some configurations by increasing recall for the disagreement class, particularly in RoBERTa-based models. However, these gains often come at the expense of majority-class performance, resulting in largely unchanged macro-F1 scores. Similarly, consistency and complementarity-based aggregation strategies yield stable but limited improvements, suggesting that while cross-modal agreement signals are informative, they are insufficient on their own to robustly capture disagreement semantics.

In contrast, the MLP-based contrastive learning approach demonstrates more consistent gains on the MAMI dataset. The contrastive MLP trained on CLIP embeddings achieves the

Method	Aggregation	Models	P-	P+	R-	R+	F1-	F1+	F1-macro
C-Fusion	Concatenation	BERT+ResNet50	0.68	0.43	0.76	0.34	0.72	0.38	**0.55***†
		BERT+ViT	0.67	0.40	0.71	0.32	0.69	0.38	0.54†
		RoBERTa+ResNet50	0.67	0.40	0.75	0.30	0.71	0.34	0.53
		RoBERTa+ViT	0.67	0.40	0.73	0.33	0.70	0.36	0.53†
	Hadamard Product	BERT+ResNet50	0.67	0.36	0.63	**0.42**	0.65	0.40	0.53
		BERT+ViT	0.67	0.43	**0.81**	0.27	0.73	0.33	0.53
		RoBERTa+ResNet50	0.68	0.39	0.65	**0.42**	0.66	**0.41**	0.54*†
		RoBERTa+ViT	0.66	0.38	0.78	0.25	0.71	0.31	0.51
	Consistency & Complementarity	BERT+ResNet50	0.67	0.38	0.65	0.40	0.66	0.39	0.53
		BERT+ViT	0.67	0.42	0.79	0.28	0.72	0.33	0.53†
		RoBERTa+ResNet50	0.66	0.38	0.64	0.40	0.65	0.39	0.52
		RoBERTa+ViT	0.68	0.40	0.74	0.32	0.71	0.35	0.53
C-MLP	—	CLIP	**0.70**	**0.45**	0.75	**0.40**	0.72	**0.42**	**0.57***†
	—	BLIP	0.69	0.44	**0.77**	0.35	0.73	0.39	0.56*†
Finetuning	—	CLIP	0.66	0.42	0.72	0.35	0.68	0.36	0.52
	—	BLIP	0.64	0.40	0.72	0.33	0.67	0.34	0.51
Rizzi et al. [125]	—	MTT	0.76	0.00	1.00	0.00	0.87	0.00	0.43

Table 6.3: Model performance on **Disagreement Detection** task on **MAMI** Dataset using 10-fold cross-validation on training data. Methods are grouped into *C-Fusion*, *C-MLP*, and supervised baselines (Finetuning, Rizzi et al. [125]). Within each Method block, **bold** values indicate the best results that outperform all corresponding baseline models, while **bold and underlined** values denote the best *F1-macro* score within that method. Symbols (*) and (†) indicate statistically significant improvements over Finetuned CLIP and BLIP, respectively.

best overall performance, with a macro-F1 score of 0.57, significantly outperforming both finetuned baselines and all fusion-based configurations. This model achieves improved precision and F1-score for the disagreement class while maintaining strong performance on the agreement class, indicating that contrastive optimization in a shared multimodal embedding space improves class separability even under severe imbalance. A similar trend is observed for the BLIP-based MLP model, which also outperforms fusion-based methods despite slightly lower recall for the disagreement class.

Notably, the MTT model exhibits degenerate behavior on MAMI, achieving perfect recall for the non-disagreement class but failing entirely to detect disagreement instances. This highlights the importance of balanced optimization objectives and further underscores the advantage of contrastive learning, which explicitly enforces both inter-class separation and intra-class cohesion.

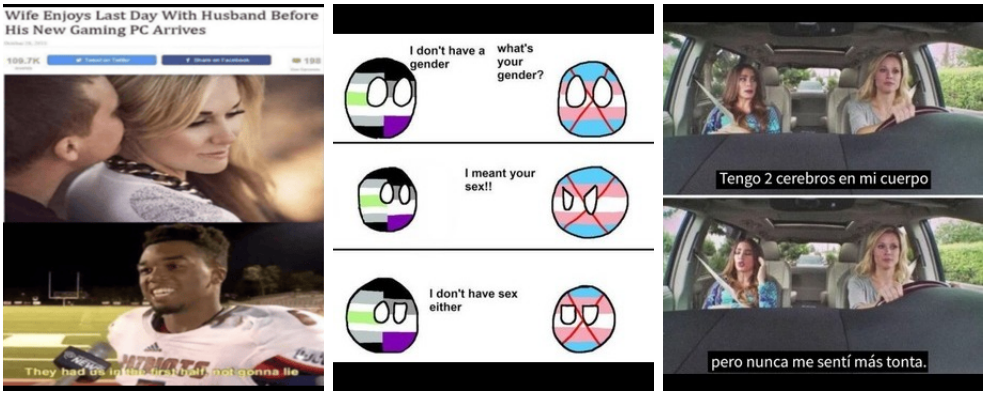
Method	Aggregation	Models	P-	P+	R-	R+	F1-	F1+	F1-macro
C-Fusion	Concatenation	BERT+ResNet50	0.78	0.30	0.84	0.22	0.81	**0.25**	**<u>0.53</u>***†
		BERT+ViT	0.77	0.29	0.84	0.20	0.80	0.23	0.52†
		RoBERTa+ResNet50	0.77	0.26	0.83	0.19	0.80	0.22	0.51
		RoBERTa+ViT	0.77	0.27	0.86	0.16	0.81	0.20	0.51
	Hadamard Product	BERT+ResNet50	0.77	0.27	0.80	0.23	0.79	0.24	0.52†
		BERT+ViT	0.77	0.29	0.88	0.16	0.82	0.20	0.51
		RoBERTa+ResNet50	0.76	0.25	0.79	0.22	0.78	0.23	0.51
		RoBERTa+ViT	0.76	0.24	0.83	0.18	0.80	0.20	0.50
	Consistency & Complementarity	BERT+ResNet50	0.77	0.28	0.81	0.24	0.79	**0.25**	0.52†
		BERT+ViT	0.76	0.26	0.88	0.14	0.82	0.18	0.50
		RoBERTa+ResNet50	**0.79**	0.28	0.80	**0.26**	0.79	0.27	**<u>0.53</u>***†
		RoBERTa+ViT	0.77	0.28	0.87	0.16	0.82	0.20	0.51
C-MLP	—	CLIP	0.78	**0.41**	0.91	0.18	**0.84**	0.25	**<u>0.55</u>***†
	—	BLIP	0.78	0.34	0.84	**0.25**	0.80	**0.29**	**<u>0.55</u>***†
Finetuning	—	CLIP	0.78	0.25	0.82	0.20	0.80	0.22	0.51
	—	BLIP	0.76	0.27	0.88	0.14	0.82	0.18	0.50
Rizzi et al. [125]	—	MTT	0.67	0.55	0.93	0.16	0.78	0.24	0.51

Table 6.4: Model performance on **Disagreement Detection** task on **EXIST** Dataset using 10-fold cross-validation on training data. Within each Method block, **bold** values indicate the best results that outperform all corresponding baseline models, while **<u>bold and underlined</u>** values denote the best *F1-macro* score within that method. Symbols (*) and (†) indicate statistically significant improvements over Finetuned CLIP and BLIP, respectively.

On the **EXIST dataset**, overall performance trends remain consistent but reflect increased robustness due to larger data diversity. Fusion-based models achieve macro-F1 scores around 0.50–0.53, with the strongest results obtained using concatenation and consistency & complementarity aggregation strategies. The RoBERTa + ResNet50 configuration with consistency and complementarity achieves a macro-F1 score of 0.53, matching the best fusion performance and showing statistically significant improvement over fine-tuned baselines.

The contrastive MLP approach again yields the best overall performance on EXIST. Both CLIP and BLIP based MLP models achieve a macro-F1 score of 0.55, significantly outperforming fine-tuned vision–language models and all fusion-based counterparts. While recall for the disagreement class remains low across all models, the contrastive MLP demonstrates improved precision and more stable minority-class F1 scores, indicating better discrimination of disagreement cues rather than indiscriminate majority-class prediction.

Across both datasets, fine-tuning pretrained vision–language models without contrastive supervision consistently underperforms compared to contrastive learning-based approaches. These findings suggest that disagreement detection benefits from explicitly structuring the



(a) Meme from MAMI dataset having True label "Disagreement" on misogyny but misclassified by all models

(b) Meme from EXIST dataset having True label "Agreement" on misogyny but misclassified by all models

(c) Meme from EXIST labeled "Agreement" correctly classified only by RoBERTa+ResNet50 (consistency & complementarity)

Figure 6.5: Illustration of Memes on Disagreement Detection

embedding space to reflect relational similarity between samples, rather than relying solely on classification loss. While explicit multimodal fusion provides competitive baselines, its dependence on handcrafted aggregation limits its ability to generalize across datasets with varying disagreement expressions.

Overall, the results demonstrate that contrastive learning offers a principled and effective mechanism for modeling disagreement in multimodal memes. Although disagreement detection remains a challenging task due to class imbalance and semantic subtlety, the proposed contrastive framework consistently improves robustness, minority-class sensitivity, and cross-dataset generalization compared to both fusion-based and fine-tuned baselines.

Figure 6.5 highlights challenges in detecting annotator disagreement on misogyny in memes and demonstrates the performance of models in addressing these complexities. Figure 6.5(a) has true label of "disagreement" regarding misogyny. However, all models misclassify it as agreement. This example illustrates the difficulty in identifying significant differences in annotator perspectives, where the ambiguity in meme content plays an important role.

Figure 6.5(b) shows the meme from EXIST dataset that has also been misclassified by all models. Its true label was "agreement" but it has been misclassified as disagreement. Figure 6.5(c) is correctly classified only by the proposed model, which utilizes RoBERTa, ViT encoders and the consistency and complementarity aggregation function. This result shows the effectiveness of the proposed approach in handling annotator agreement.

6.3.3 Cross-Task Analysis

Having presented results for both hate (misogyny/sexism) detection (Section 6.3.1) and disagreement detection (Section 6.3.2), we now examine what these results reveal when considered together.

Shared Representational Foundation. The same architecture and pretrained encoder (C-MLP with CLIP) achieves the best performance on both hate detection (F1-macro 0.76) and disagreement detection (F1-macro 0.57). This consistency suggests that CLIP's pretrained vision-language representations encode information relevant to both identifying misogynistic content and recognizing content that will elicit human perceptual disagreement. Moreover, the MLP-on-embeddings approach consistently outperforms fusion architectures on both tasks (0.76 vs 0.72 for hate; 0.57 vs 0.55 for disagreement), indicating that operating directly on joint pretrained embeddings better preserves the vision-language alignment crucial for both phenomena.

The Difficulty Gap and Its Meaning. Disagreement detection is substantially harder than hate detection, with even the best method achieving F1-macro of 0.57 compared to 0.76 for hate. This performance gap reflects a fundamental difference in task complexity. Hate detection requires recognizing patterns that consistently correlate with misogynistic content. Disagreement detection, by contrast, requires predicting when those very patterns will be ambiguous to humans: a meta-level task where the model must learn not just what features indicate hate, but what features will lead some annotators to perceive hate while others do not. This second-order complexity explains why disagreement detection remains challenging even with powerful representations.

SOTA Failure on Disagreement. The state-of-the-art baseline (Rizzi et al.) achieves competitive hate detection performance (F1-macro 0.71) but completely fails on disagreement detection (F1+ = 0.00), collapsing to majority-class prediction. This stark contrast reveals that methods optimized exclusively for hate detection develop no capacity to recognize ambiguous content. Our contrastive framework, by contrast, maintains reasonable performance on both tasks, suggesting that the representational learning it induces is more transferable across related subjective classification problems.

Architectural Consistency Across Tasks. The consistent advantage of the MLP-on-embeddings approach across both tasks provides architectural guidance: fusion-based methods that separately encode modalities before combining them appear to disrupt the vision-language alignment learned during pretraining. Operating directly on joint embeddings preserves this alignment, which our results suggest is critical for tasks involving semantic ambiguity and subjective judgment.

6.3.4 Experimental Comparison: PaG-SCon *vs* SupCon

To further evaluate the performance of proposed PaG-SCon, we compare the best configuration from both fusion-based and MLP-based approaches to the same models trained with the original SupCon loss. The results are summarized in Table 6.5 for both misogyny and disagreement detection tasks on the MAMI and EXIST datasets.

Overall, PaG-SCon consistently matches or improves upon SupCon across most settings, with clear gains observed in macro-F1 scores. For misogyny detection on MAMI, PaG-SCon improves macro-F1 for both fusion-based and MLP-based models by enhancing class balance, particularly through higher recall and F1 scores for the non-misogynous class. On the EXIST dataset, statistically significant improvements are observed for fusion-based models, while performance for MLP-based CLIP models remains stable, indicating that strong pretrained multimodal alignment already captures much of the task structure.

For disagreement detection, the benefits of PaG-SCon are more noticeable in minority-class recognition. On both MAMI and EXIST, PaG-SCon improves macro-F1 by increasing recall and F1-score for the disagreement class, with the largest and statistically significant gains observed for the MLP-based CLIP model on EXIST. These results indicate that PaG-SCon often leads to better macro performance and improves the ability of the model to generalize across classes, especially in imbalanced settings.

Additionally, it is important to note that the proposed PaG-SCon can have a reduced computational complexity compared to the SupCon. In particular, in the SupCon loss we need to compute for each anchor z_m the dot product with all other embeddings $z_a \neq z_m$. Considering N elements in the batch and V as embedding size, the SupCon complexity is $O(N^2V)$. On the other hand, in the PaG-SCon loss we focus our attention on intra-class paired samples, according to the definition of $\mathcal{P}$, making this loss more computationally efficient than the original SupCon or in the worst case analogous. In particular, by defining $|\mathcal{P}| = \{\langle z_m, z_k \rangle \mid k < m \text{ and } y_m = y_k\}$ as the set of all pairs belonging to the same class, we enforce intra-class cohesion encouraging semantic consistency within the same class. This settings allows us to consider intra-class pairs $\mathcal{P} = \bigcup_{c=1}^{C} \binom{n_c}{2}$, where n_c is the number of samples belonging to class $c \in C$.

Now, let's consider two opposite scenarios, i.e., a balanced case, where $n_c = \frac{N}{C}$, and the worst scenario with a strong imbalance between classes (e.g. one dominant class containing almost all samples in the batch). In the first case, PaG-SCon requires only $O\left(\frac{N^2}{C}\right)$ dot products, while in the worst scenario corresponds to the original SupCon that requires $O(N^2)$ estimates.

Task	Dataset	Method	Models	Loss Function	P-	P+	R-	R+	F1-	F1+	F1-macro
Misogyny	MAMI	C-Fusion	BERT+ViT	SupCon	0.81	0.65	0.51	0.90	0.63	0.75	0.70
			(Hadamard Product)	PaG-SCon	0.81	**0.68**	**0.59**	0.87	**0.68**	**0.76**	**0.72**
		C-MLP	CLIP	SupCon	0.91	0.68	0.56	0.94	0.69	0.79	0.74
				PaG-SCon	0.88	**0.70**	**0.61**	0.92	**0.72**	**0.80**	**0.76**
Sexism	EXIST	C-Fusion	BERT+ResNet50	SupCon	0.62	0.71	0.58	0.74	0.59	0.73	0.66
			(Concatenation)	PaG-SCon	**0.68**	0.70	0.55	**0.79**	**0.61**	**0.74**	**0.68***
		C-MLP	CLIP	SupCon	0.62	0.75	0.49	0.84	0.54	0.79	0.67
				PaG-SCon	0.61	0.75	0.48	0.84	0.54	0.79	0.67
Disagreement	MAMI	C-Fusion	BERT+ResNet50	SupCon	0.68	0.42	0.79	0.30	0.72	0.35	0.54
			(Concatenation)	PaG-SCon	**0.68**	**0.43**	0.76	**0.34**	**0.72**	**0.38**	**0.55***
		C-MLP	CLIP	SupCon	0.70	0.47	0.81	0.32	0.75	0.38	0.56
				PaG-SCon	0.70	0.45	0.75	**0.40**	0.72	**0.42**	**0.57**
Disagreement	EXIST	C-Fusion	RoBERTa+ResNet50	SupCon	0.78	0.31	0.86	0.18	0.83	0.23	0.53
			(Consistency & Complementarity)	PaG-SCon	**0.79**	0.28	0.80	**0.26**	0.79	**0.27**	0.53
		C-MLP	CLIP	SupCon	0.77	0.35	0.91	0.17	0.84	0.22	0.53
				PaG-SCon	**0.78**	**0.41**	0.91	**0.18**	0.84	**0.25**	**0.55***

Table 6.5: Model performance on **Misogyny/Sexism** and **Disagreement** tasks across the MAMI and EXIST datasets using SupCon and PaG-SCon loss functions. For each method, **bold** values indicate metrics that outperform the corresponding baseline models. **Bold and underlined** values denote the best *F1-macro* score achieved. Entries marked with (*) indicate statistically significant improvements over SupCon.

6.3.5 Environmental Impact Analysis

To complement the performance evaluation, we assessed the environmental impact of each method by estimating their carbon footprint using the metrics described in Section 4.3.2. This analysis is particularly relevant in the context of *Green AI*, where the objective is to design models that achieve competitive performance while minimizing computational cost and energy consumption.

The results reveal a clear distinction between the proposed contrastive learning approaches (C-Fusion and C-MLP) and the finetuned vision–language baselines. Across all tasks and datasets, the contrastive learning models, specially the lightweight **C-MLP**, achieve the lowest CO_2 emissions, often by a substantial margin. Importantly, this efficiency does not compromise predictive performance. In several settings, C-MLP not only matches but **outperforms** both C-Fusion and the computationally intensive finetuned models.

A key factor contributing to these differences is the GPU utilization required by each method. Finetuning the vision–language models consistently consumed **80–90%** of the available GPU capacity, whereas C-Fusion required approximately **70–80%**, and C-MLP operated at only **40–50%**. Higher GPU utilization directly translates into increased energy consumption, which in turn results in higher carbon emissions. This relationship is clearly reflected in the estimated CO_2 values.

Task	Dataset	Method	Models	F1-macro	Exec Time	gCO_2
Misogyny	MAMI	C-Fusion	BERT+ViT (HP)	0.72	83.9	53.0
		C-MLP	CLIP	**0.76**	**8.64**	**3.64**
		Finetuning	CLIP	0.70	73.7	55.8
Sexism	EXIST	C-Fusion	BERT+ResNet50 (Conc)	**0.68**	11.8	7.42
		C-MLP	CLIP	0.67	**3.44**	**1.45**
		Finetuning	CLIP	17.51	13.3	10.1
Disagreement	MAMI	C-Fusion	BERT+ResNet50 (Conca)	0.55	30.9	19.5
		C-MLP	CLIP	**0.57**	**6.87**	**2.89**
		Finetuning	CLIP	0.52	75.1	56.9
Disagreement	EXIST	C-Fusion	RoBERTa+ResNet50 (C&C)	0.53	14.36	9.10
		C-MLP	CLIP	**0.55**	**3.13**	**1.32**
		Finetuning	CLIP	0.51	14.2	10.8

Table 6.6: Environmental impact analysis of different multimodal methods across tasks and datasets. **Bold** values indicate the best F1-macro performance within each task–dataset setting, while **bold and underlined** values highlight the method achieving the lowest computational cost in terms of execution time and CO_2 emissions.

For instance, on the MAMI misogyny task, C-MLP achieves the highest F1-macro score (0.76) while emitting only **3.64 gCO_2**, compared to **53.0 gCO_2** for C-Fusion and **55.8 gCO_2** for the finetuned CLIP model. Similar trends appear in the disagreement detection tasks, where C-MLP again delivers the best performance with emissions as low as **2.89 gCO_2** on MAMI and **1.32 gCO_2** on EXIST. These values represent an order-of-magnitude reduction in carbon cost relative to the finetuned vision–language models.

In contrast, the finetuned models consistently exhibit the highest energy consumption despite not always achieving superior performance. Their execution times are significantly longer, and their estimated emissions are up to **20–40 times higher** than those of C-MLP. This highlights a key limitation of large vision–language architectures: while powerful, they are often impractical for real-time or large-scale moderation systems due to their environmental and computational overhead.

Overall, the findings demonstrate that the proposed contrastive learning framework provides a highly effective and sustainable alternative to traditional finetuning. In particular, C-MLP offers a compelling balance between accuracy, efficiency, and environmental responsibility, making it well suited for deployment in real-world content moderation pipelines where scalability and sustainability are critical considerations.

6.4 Extending Contrastive Learning to Joint Hate–Disagreement Detection

In addition to binary hate (misogyny/sexism) detection, we extend the proposed contrastive learning framework to the more challenging task of joint hate and disagreement detection, where each meme expresses both a stance (agreement or disagreement) and a harmful intent (hateful or non-hateful). This task is inherently more complex, as it requires the model to capture interdependent semantic attributes rather than a single label dimension.

To analyze the interaction between misogyny and annotator, we created a joint confusion matrix using the best C-Fusion model on the EXIST dataset from the previous section (see Figure 6.6). Specifically, the joint confusion matrix reflects the distribution over four possible combinations: *Agreement–Not Hateful, Agreement–Hateful, Disagreement–Not Hateful, and Disagreement–Hateful*. It provides a holistic view of how the predictions of the model align with the true labels when considering both dimensions simultaneously.

The confusion matrix allows us to observe that a significant number of memes labeled as *Agreement-Hateful* were misclassified as *Disagreement-Hateful*. Similarly, memes labeled as *Disagreement-Hateful* were frequently misclassified as *Agreement-Hateful*. This indicates a difficulty in distinguishing between disagreement and agreement when hateful content is present. We also observe that the model generally classified *Disagreement-Hateful* instances correctly, indicating that it has learned some discriminative features specific to this class.

To study the effectiveness of contrastive learning in this setting, we conduct experiments under two complementary formulations: **multi-class classification** and **multi-label classification**. In both cases, we use the same image–text embedding backbone and contrastive learning objective, with modifications only at the label modeling level.

6.4.1 Multi-Class Joint Classification

In the first setting, joint hate–disagreement detection is formulated as a multi-class classification problem, where each meme is assigned exactly one of four mutually exclusive classes. These classes correspond to all possible combinations of stance and hate labels:

- Agreement–Non-Hate (0)
- Disagreement–Non-Hate (1)
- Agreement–Hate (2)
- Disagreement–Hateful (3)

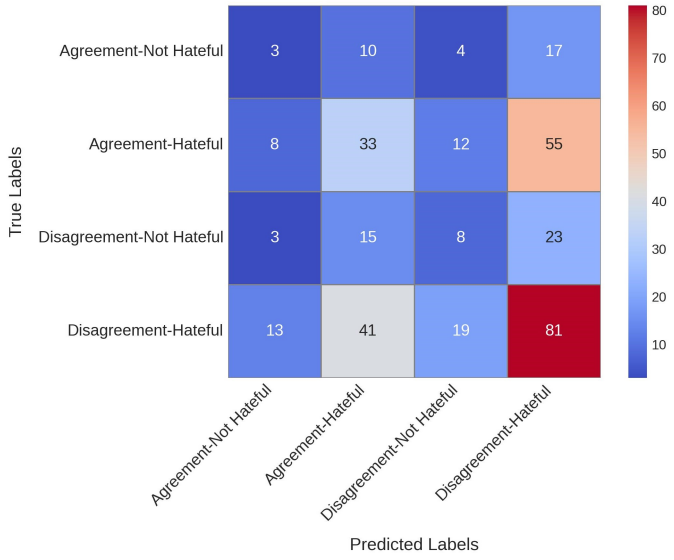


Figure 6.6: Confusion matrix showing the interaction between misogyny and disagreement for EXIST dataset

Under this formulation, the contrastive learning framework remains unchanged. Each meme is represented by a joint image–text embedding z_i, and supervised contrastive learning is applied by treating samples that share the same joint class label as positive pairs, while all other pairs are treated as negatives.

Formally, given a batch of embeddings $\{z_1, z_2, \ldots, z_n\}$ with corresponding joint labels $y_i \in \{0, 1, 2, 3\}$, the set of positive pairs is defined as:

$$\langle z_m, z_k \rangle \in \mathcal{P} \quad \text{if and only if} \quad y_m = y_k.$$

This formulation enforces compact clustering of samples belonging to the same joint semantic category while maximizing separation across different categories. Although this approach provides a strict and unambiguous supervision signal, it does not explicitly model partial semantic overlap between classes (e.g., between Agreement–Hateful and Agreement–Non-Hateful).

Labels		11	10	00	11	01
		z1	**z2**	**z3**	**z4**	**z5**
11	**z1**	1	--	--	1	--
10	**z2**	--	1	--	--	--
00	**z3**	--	--	1	--	--
11	**z4**	1	--	--	1	--
01	**z5**	--	--	--	--	1

(a) Example of Contrastive learning matrix for a multiclass setup, where each sample has a single positive region.

Labels		11	10	00	11	01
		z1	**z2**	**z3**	**z4**	**z5**
11	**z1**	1	0.5	0	1	0.5
10	**z2**	--	1	0.5	0.5	0.5
00	**z3**	--	--	1	0	0.5
11	**z4**	1	--	--	1	0.5
01	**z5**	--	--	--	--	1

(b) Example of Contrastive learning matrix for a multilabel setup, where multiple labels create multiple positive regions.

Figure 6.7: Comparison of contrastive learning matrices under multiclass and multilabel supervision.

6.4.2 Multi-Label Joint Classification

In the second setting, joint detection is formulated as a *multi-label classification* problem, where hate and disagreement are treated as two correlated but distinct binary labels. Each meme is associated with a label vector:

$$\mathbf{y}_i = \left(y_i^{\text{hate}},\ y_i^{\text{disagreement}}\right),$$

where each component takes a value in $\{0, 1\}$.

Unlike the multi-class formulation, this setting allows samples to share partial label similarity. For example, two memes may both express disagreement but differ in hateful intent. Given that Disagreement/Hate is denoted as 1 and Agreement/Non-Hate is denoted as 0. We have following combinations:

- Disagreement-Hate (11)
- Disagreement-NonHate (10)
- Agreement-Hate (01)
- Agreement-NonHate (00)

To incorporate this structure into representation learning, the contrastive objective is extended with a *label-aware similarity weighting*. For any pair of samples (i, j), a weight

$$w_{ij} = \frac{|y_i \cap y_j|}{|y_i \cup y_j|}$$

is computed using the Jaccard similarity between their label vectors. Pairs sharing both labels receive the highest weight, pairs sharing exactly one label receive an intermediate weight, and pairs with no shared labels are treated as negatives. This graded weighting encourages the model to pull semantically related samples closer in the embedding space while maintaining separation across unrelated instances, consistent with established practices in multi-label contrastive learning.

To assess how well the learned embedding space preserves this multi-label structure, I additionally employ a *Jaccard-weighted* k-nearest neighbors (k-NN) classifier during evaluation. For each test embedding, the top-k nearest training embeddings are retrieved using cosine similarity. Unlike the training stage—where weights are derived from label overlap—here the weighting is computed directly from the embedding geometry: a Jaccard-style score based on the intersection–union ratio of the embedding vectors. These continuous weights are then used to perform weighted voting over the two binary labels, producing a prediction vector

$$\left(\hat{y}^{\text{hate}},\ \hat{y}^{\text{disagreement}}\right).$$

This non-parametric evaluation provides an interpretable measure of whether the learned representation space meaningfully reflects the underlying multi-label relationships.

6.4.3 Experimental Settings

All experiments were conducted on the **EXIST** and **MAMI** datasets following the 10-fold cross-validation. An analysis of the joint hate and disagreement label distributions revealed a substantial class imbalance, particularly for agreement-related categories, which poses additional challenges for joint modeling. The distribution of the labels for both datasets is shown in Figures 6.8 and 6.9.

For **Fusion-based approach**, we selected the ViT + BERT configuration with Hadamard Product aggregation strategy. This choice was motivated by preliminary experiments, where the Hadamard Product consistently yielded relatively better performance across the individual misogyny and disagreement detection tasks.

For the **MLP-based approach**, we employed CLIP to extract pretrained multimodal embeddings, followed by a lightweight MLP projection network. CLIP was chosen because it demonstrated the most robust and consistent performance for both misogyny and disagreement detection among the evaluated vision–language models. The projected embeddings were subsequently optimized using the contrastive learning objective described in Section 2.4.

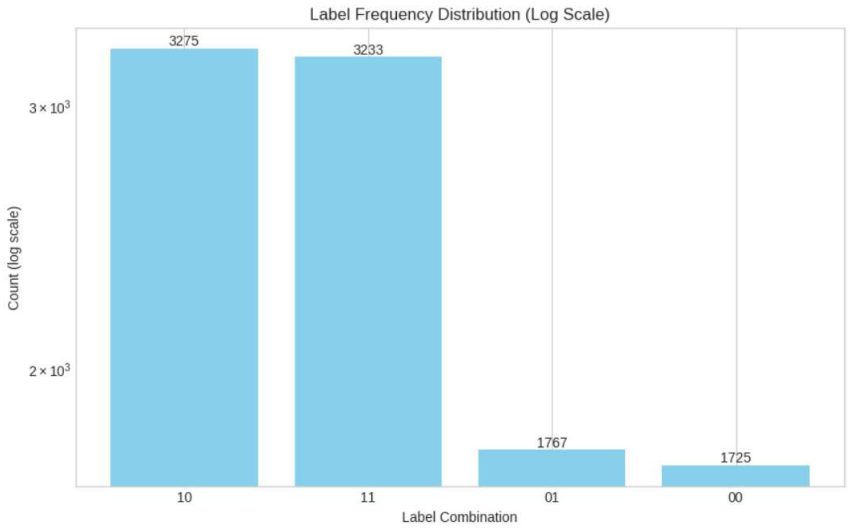


Figure 6.8: Label distribution matrix showing the interaction between misogyny and disagreement for the **MAMI** dataset.

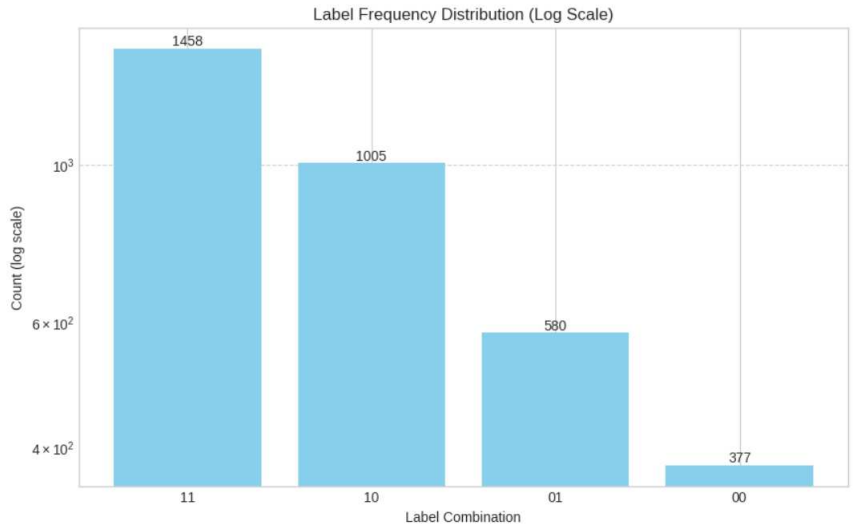


Figure 6.9: Label distribution matrix showing the interaction between misogyny and disagreement for the **EXIST** dataset.

Method	Class	Precision	Recall	F1-score
C-Fusion	Agreement-NonHate (0)	0.21	0.08	0.12
	Disagreement-NonHate (1)	0.28	0.24	0.26
	Agreement-Hate (2)	0.23	0.12	0.16
	Disagreement-Hate (3)	0.53	0.71	0.61
	Macro Avg	0.29	0.37	0.26
	Weighted Avg	0.41	0.47	0.41
C-MLP	Agreement-NonHate (0)	0.37	0.24	0.29
	Disagreement-NonHate (1)	0.44	0.26	0.33
	Agreement-Hate (2)	0.40	0.18	0.24
	Disagreement-Hate (3)	0.56	0.80	0.66
	Macro Avg	0.44	0.37	**0.38**
	Weighted Avg	0.49	0.52	**0.48**

Table 6.7: Multiclass Classification Results on EXIST Dataset

6.4.4 Results and Discussion

We evaluate our proposed contrastive learning framework on the joint modeling of misogyny and annotator disagreement, formulated as both *multiclass* and *multilabel* classification problems. Experiments are conducted on the **EXIST** and **MAMI** datasets using two embedding construction strategies: **C-Fusion** and **C-MLP**. We note that both datasets exhibit noticeable class imbalance, particularly for agreement-related classes, which influences overall performance trends.

Multiclass Classification

The multiclass classification results are presented in Tables 6.7 and 6.8, where each meme is assigned to one of four mutually exclusive classes that represent the combined states of hate and disagreement.

On the **EXIST** dataset, **C-MLP** consistently outperforms **C-Fusion**, achieving a higher macro F1-score (0.38 vs. 0.26) and weighted F1-score (0.48 vs. 0.41). Performance gains are particularly evident for the *Agreement–NonHate* and *Disagreement–NonHate* classes, suggesting that the MLP-based projection is effective in refining pretrained CLIP/BLIP representations for subtle semantic distinctions. Both methods achieve their strongest performance on the *Disagreement–Hate* class, with C-MLP reaching an F1-score of 0.66. This aligns with the class distribution in EXIST, where the *Disagreement–Hate* class accounts for approximately 42.6% of the data, making it the most frequently observed and hence easier to model.

Loss Function	Class	Precision	Recall	F1-score
C-Fusion	Agreement-NonHate (0)	0.30	0.17	0.22
	Disagreement-NonHate (1)	0.68	0.80	0.73
	Agreement-Hate (2)	0.30	0.18	0.22
	Disagreement-Hate (3)	0.64	0.78	0.70
	Macro Avg	0.48	0.48	0.47
	Weighted Avg	0.54	0.59	0.55
C-MLP	Agreement-NonHate (0)	0.34	0.26	0.29
	Disagreement-NonHate (1)	0.69	0.75	0.72
	Agreement-Hate (2)	0.28	0.22	0.24
	Disagreement-Hate (3)	0.62	0.72	0.67
	Macro Avg	0.48	0.49	**0.48**
	Weighted Avg	0.54	0.56	**0.54**

Table 6.8: Multiclass Classification Results on MAMI Dataset

On the **MAMI** dataset, the performance gap between the two approaches is narrower. C-MLP again achieves the best overall performance, with a macro F1-score of 0.48 and a weighted F1-score of 0.54. Both methods perform well on disagreement-related classes, particularly *Disagreement–NonHate*, which is one of the dominant classes in MAMI. In contrast, *Agreement–Hate* remains challenging for both models, reflecting both its lower prevalence and the difficulty of identifying subtle hateful content when annotator consensus is high.

Overall, multiclass results indicate that C-MLP benefits from strong pretrained multimodal representations, while C-Fusion is more affected by class imbalance and semantic overlap between agreement-related categories.

Multilabel Classification

We evaluate multilabel classification performance by jointly modeling misogyny and disagreement as correlated but independent labels, with results summarized in Tables 6.9 and 6.10.

On the **EXIST** dataset, **C-Fusion** achieves a higher macro F1-score (0.28) compared to C-MLP (0.25), while both methods obtain identical weighted F1-scores (0.38). C-Fusion demonstrates improved performance on the *Disagreement–NonHate* class, indicating that explicit fusion of unimodal embeddings better preserves complementary signals relevant to annotator disagreement. Both approaches struggle with *Agreement–Hate* and *Agreement–NonHate*, which together constitute less than 30% of the dataset, highlighting the impact of label imbalance on multilabel learning.

Loss Function	Class	Precision	Recall	F1-score
C-Fusion	Disagreement-Hate (11)	0.50	0.73	0.59
	Disagreement-NonHate (10)	0.34	0.19	0.24
	Agreement-Hate (01)	0.17	0.10	0.13
	Agreement-NonHate (00)	0.23	0.11	0.15
	Macro Avg	0.31	0.29	**0.28**
	Weighted Avg	0.38	0.42	**0.38**
C-MLP	Disagreement-Hate (11)	0.50	0.71	0.59
	Disagreement-NonHate (10)	0.31	0.27	0.29
	Agreement-Hate (01)	0.17	0.07	0.10
	Agreement-NonHate (00)	0.17	0.01	0.02
	Macro Avg	0.29	0.26	0.25
	Weighted Avg	0.37	0.43	**0.38**

Table 6.9: Multi-Label Classification Results on EXIST Dataset

Method	Class	Precision	Recall	F1-score
C-Fusion	Disagreement-Hate (11)	0.53	0.72	0.61
	Disagreement-NonHate (10)	0.63	0.74	0.68
	Agreement-Hate (01)	0.28	0.19	0.23
	Agreement-NonHate (00)	0.30	0.09	0.14
	Macro Avg	0.44	0.44	**0.42**
	Weighted Avg	0.48	0.53	**0.49**
C-MLP	Disagreement-Hate (11)	0.33	0.54	0.41
	Disagreement-NonHate (10)	0.34	0.30	0.32
	Agreement-Hate (01)	0.19	0.15	0.17
	Agreement-NonHate (00)	0.16	0.037	0.06
	Macro Avg	0.26	0.26	0.24
	Weighted Avg	0.28	0.31	0.28

Table 6.10: Multi-Label Classification Results on MAMI Dataset

On the **MAMI** dataset, C-Fusion outperforms C-MLP, achieving a macro F1-score of 0.42 compared to 0.24. The advantage is most pronounced for disagreement-related classes, which collectively form the majority of the dataset. The explicit aggregation of image and text embeddings enables C-Fusion to better capture multimodal ambiguity and conflicting cues. In contrast, the MLP-based projection appears to compress these signals, resulting in degraded performance, particularly for agreement classes.

Across both datasets and task formulations, several consistent trends emerge. First, *Disagreement-Hate* is the easiest class to identify, benefiting from both strong multimodal cues and higher class prevalence. Second, C-MLP is better suited for multiclass prediction, where a single dominant label must be inferred, while C-Fusion exhibits greater robustness in multilabel settings, especially for disagreement detection. Finally, agreement-related classes remain challenging across all experiments, largely due to their lower frequency and the subjective nature of annotator consensus. These findings highlight the importance of embedding construction strategies and suggest that future work could benefit from explicitly modeling class imbalance and annotator uncertainty.

6.5 Current limitations and impact on theoretical and practical aspects

Despite the promising results demonstrated by the proposed multimodal models in hate and disagreement detection, this study has several limitations. First, the performance variations between MAMI and EXIST datasets (e.g., optimal aggregation functions differing by dataset) highlight *potential biases* in training data composition, such as uneven representation of meme styles or annotator perspectives. The proposed approach treats all relationships equally, which might weaken its ability to model fine-grained intra-class variations. Second, although our models show strong performance in detecting misogyny and disagreement, they occasionally fail to distinguish between *nuanced cases*, particularly when hateful content lead to disagreement. This suggests limitations in the ability of model to capture subtle semantic or contextual cues.

The present study also has several theoretical and practical impacts. Our method shifts from relying on anchor-based comparisons to using a global normalization over all sample pairs in the batch. This change helps the model learn similarity relationships more effectively. By treating all positive pairs equally, we avoid problems that come from having different numbers of positive samples per anchor. This introduces a form of structural regularization, as embeddings are encouraged to form compact class-specific subspaces. An additional theoretical implication relates to computational complexity. In particular, our proposed approach also reduces the number of dot product calculations, which helps speed up training especially when working with balanced or slightly imbalanced datasets.

This research has several useful practical implications. First, our models can make automated content moderation more effective by better spotting misogynistic content, particularly in multimodal formats such as memes, where meaning often arises from the interplay between text and image. Second, by predicting disagreement, the model captures uncertainty in

how content is interpreted. This allows it to identify unclear or borderline cases early, which can then be reviewed by human moderators. As a result, errors are reduced and decision-making becomes more accurate and consistent. Finally, the proposed contrastive loss function demonstrates promising performance across complex and subjective tasks, encouraging the development of models used in sensitive areas like hate speech detection.

Future work can explore several directions to further improve the current approach. An ongoing research regards the definition of an adaptive weighting strategy for positive pairs, allowing the model to emphasize more informative or difficult samples. This could lead to better generalization, especially in scenarios where subtle contents are present. Additionally, extending the loss to handle multi-label or hierarchical class structures could enhance its effectiveness in more complex classification tasks, such as the recognition of the misogyny category (e.g. stereotype, violence against women), to better reflect the layered and nuanced nature of harmful communication.

Chapter Summary

In this chapter, we proposed a contrastive learning approach to identify misogynous memes and model perceptual disagreement among human annotators. Our experiments evaluate various combinations of image and text encoders alongside distinct aggregation functions to assess their impact on performance. We further utilized embeddings from vision–language models to train a lightweight MLP under the same contrastive objective. In addition, we extended the framework to the joint detection of hate and disagreement, demonstrating that the proposed formulation can generalize beyond binary labels to capture richer, multi-attribute semantics. Overall, the results show that our approach outperforms current state-of-the-art methods, highlighting its effectiveness in both misogyny detection and disagreement analysis.

7 Conclusion and Future Work

This thesis investigated the problem of sexism and misogyny detection in multimodal online content, with a particular focus on memes, while also addressing the crucial aspect of perceptual disagreement among human annotators. Motivated by the rapid growth of harmful user-generated content on social media and the limitations of purely human moderation, this work aimed to design automated methods that are not only effective but also computationally efficient and environmentally sustainable.

One of the central challenges highlighted in the thesis is the inherent subjectivity involved in identifying hate and offensive content. Perceptions of misogyny and sexism vary widely across individuals and cultural contexts, leading to disagreement that reflects the complexity of human judgment rather than annotation noise. By explicitly acknowledging and modeling this variability, the thesis moves beyond traditional binary classification and contributes to a more realistic and nuanced understanding of harmful content detection.

To address these challenges, the thesis proposed lightweight multimodal frameworks based on pretrained representations, avoiding the heavy computational costs associated with large-scale multimodal models. First, a semi-supervised constrained clustering approach was introduced, leveraging embedding regularization to exploit the representational power of pretrained vision–language models while reducing the need for fine-tuning and labeled data. Although this approach achieved slightly lower performance than fully supervised methods, it demonstrated substantial reductions in computational cost and carbon energy consumption, emphasizing the feasibility of greener alternatives for real-world deployment.

Second, a contrastive learning–based framework was proposed to detect misogynistic content and annotator disagreement. By directly optimizing similarities between label-consistent embedding pairs, the method effectively captured both semantic alignment and perceptual variability. Extensive experiments across different encoder combinations, fusion strategies, and embedding sources showed that the approach consistently outperformed state-of-the-art baselines. Furthermore, the extension to joint hate–disagreement detection demonstrated the flexibility of the formulation in handling richer, multi-attribute classification tasks.

A key contribution of this thesis lies in its systematic analysis of the trade-off between performance and energy consumption. The results show that competitive accuracy can be

achieved without resorting to resource-intensive models, reinforcing the relevance of Green AI principles in socially responsible machine learning research. By quantifying environmental costs alongside predictive performance, this work provides practical insights for the development of scalable and sustainable content moderation systems.

Overall, this thesis advances the state of the art in multimodal misogyny detection by integrating efficiency, sustainability, and subjectivity-aware modeling. The proposed methods offer a promising direction for building moderation tools that are not only accurate but also ethically and environmentally conscious, addressing both the social impact of harmful online content and the ecological footprint of modern AI systems.

Future Research Directions

While this thesis makes several contributions to multimodal misogyny/sexism and disagreement detection, it also opens up multiple avenues for future research.

- One promising direction concerns the refinement of the contrastive learning framework. In particular, adaptive weighting strategies for positive pairs could be explored, allowing the model to place greater emphasis on informative, ambiguous, or difficult samples. Such strategies may improve generalization, especially in cases where misogyny is subtle, implicit, or heavily context-dependent. Incorporating curriculum-based or uncertainty-aware weighting schemes could further enhance robustness.

- Another promising direction for future research lies in extending the proposed semi-supervised clustering framework through the introduction of a dynamic penalty term, replacing the current fixed regularization strategy. Rather than enforcing uniform constraint strength across all samples and training stages, an adaptive penalty could vary according to contextual factors such as clustering stability, sample ambiguity, local embedding density, or annotator agreement. Such adaptive mechanisms have the potential to improve clustering quality, better reflect the subjective nature of hate speech annotation, and further close the performance gap with fully supervised state-of-the-art models while maintaining computational efficiency.

- A further important extension involves moving beyond binary labels toward multi-label and hierarchical classification. Misogyny manifests in diverse forms, such as stereotypes, objectification, harassment, and violence against women. Extending the proposed methods to capture these fine-grained categories would better reflect the layered nature of harmful communication and enable more precise moderation and analysis.

Hierarchical contrastive objectives or structured label representations could be particularly effective in this context.

- From a multimodal perspective, additional research could investigate the role of cross-modal interactions at a finer granularity, such as region-level visual features aligned with specific textual spans. While maintaining computational efficiency, lightweight attention or alignment mechanisms may help capture subtle visual cues that contribute to misogynistic meaning in memes.

- Finally, expanding the evaluation to more diverse languages, cultural contexts, and platforms represents an important future direction. Since perceptions of sexism and misogyny are strongly influenced by socio-cultural factors, testing the proposed approaches in multilingual and cross-cultural settings would enhance their generalizability and practical relevance. Combining this with continued analysis of energy consumption across different deployment scenarios would further support the development of truly sustainable and inclusive content moderation systems.

Beyond individual methodological extensions, the long-term goal of this research is to move toward content moderation systems that are essentially adaptive, and consistent with human values. Future research should focus on modeling hateful communication as a dynamic, context-dependent phenomenon impacted by cultural norms, rather than considering misogyny and sexism detection as a static prediction problem. This includes developing moderation frameworks that can represent uncertainty and disagreement and adapt over time as societal definitions of harm change. In parallel, the emphasis on computational efficiency and environmental sustainability will be essential to ensure that such systems remain deployable at scale without increasing ecological or social problems. Ultimately, the long-term objective is to promote responsible AI-driven moderation that respects cultural diversity, enhances human judgment and makes online environments safer and more inclusive while remaining mindful of the environmental costs of large-scale AI deployment.

In summary, future research can build upon the foundations laid in this thesis by deepening the modeling of subjectivity, expanding semantic granularity, and further balancing performance with sustainability, ultimately contributing to more responsible and effective AI-driven solutions for combating online misogyny and sexism.

Bibliography

[1] Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. Semi-supervised multi-task learning for multi-label fine-grained sexism classification. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5810–5820. International Committee on Computational Linguistics, 2020.

[2] Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. *Data Sci. Eng.*, 6(4):359–379, 2021.

[3] Stephen Akuma, Tyosar Lubem, and Isaac Terngu Adom. Comparing bag of words and TF-IDF with different models for hate speech detection from live tweets. *International Journal of Information Technology*, 14(7):3629–3635, 2022.

[4] Areej Al-Hassan and Hmood Al-Dossari. Detection of hate speech in arabic tweets using deep learning. *Multimedia Systems*, 28(6):1963–1974, 2022.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[6] Safa Alsafari and Samira Sadaoui. Semi-supervised self-training of hate and offensive speech from social media. *Applied Artificial Intelligence*, 35(15):1621–1645, 2021.

[7] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on twitter. In Max Silberztein, Faten Atigui,

Elena Kornyshova, Elisabeth Métais, and Farid Meziane, editors, *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, volume 10859 of *Lecture Notes in Computer Science*, pages 57–64. Springer, 2018.

[8] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on twitter. In Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane, editors, *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer, 2018.

[9] Stavros Assimakopoulos, Fabienne H. Baider, and Sharon Millar. *Online Hate Speech in the European Union*. Springer Cham, 2017.

[10] Eniafe Festus Ayetiran and Özlem Özgöbek. An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection. *Information Systems*, 123:102378, 2024.

[11] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, and Adebayo Abayomi-Alli. A probabilistic clustering model for hate speech classification in twitter. *Expert Syst. Appl.*, 173:114762, 2021.

[12] Thong Bach, Anh Tong, Truong Son Hy, Vu Nguyen, and Thanh Nguyen-Tang. Global contrastive learning for long-tailed classification. *Trans. Mach. Learn. Res.*, 2023, 2023.

[13] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019.

[14] Deeparghya Dutta Barua, MSUR Sourove, Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Fahim, and Farhad Alam Bhuiyan. Penta ml at exist 2024: tagging sexism in online multimodal content with attention-enhanced modal context. *Working Notes of CLEF*, 2024.

[15] Mequanent Degu Belete and Girma Kassa Alitasb. Identification of hateful amharic language memes on facebook using deep learning algorithms. *Systems and Soft Computing*, 7:200258, 2025.

[16] Nijole V. Benokraitis and Joe R. Feagin. *Modern sexism : blatant, subtle, and covert discrimination*. Prentice Hall, Englewood Cliffs, N.J, 2nd ed. edition, 1995.

[17] Marina Bertolini, Pierdomenico Duttilo, and Francesco Lisi. Accounting carbon emissions from electricity generation: A review and comparison of emission factor-based methods. *Applied Energy*, 392:125992, 2025.

[18] Ummara Bibi, Mahrukh Mazhar, Dilshad Sabir, Muhammad Fasih Uddin Butt, Ali Hassan, Mustansar Ali Ghazanfar, Arshad Ali Khan, and Wadood Abdul. Advances in pruning and quantization for natural language processing. *IEEE Access*, 12:139113–139128, 2024.

[19] Ummara Bibi, Mahrukh Mazhar, Dilshad Sabir, Muhammad Fasih Uddin Butt, Ali Hassan, Mustansar Ali Ghazanfar, Arshad Ali Khan, and Wadood Abdul. Advances in pruning and quantization for natural language processing. *IEEE Access*, 12:139113–139128, 2024.

[20] Verónica Bolón-Canedo, Laura Morán-Fernández, Brais Cancela, and Amparo Alonso-Betanzos. A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599:128096, 2024.

[21] Taha Bouhsine, Imad El Aaroussi, Atik Faysal, and Wang Huaxia. Simo loss: Anchor-free contrastive loss for fine-grained supervised contrastive learning. *CoRR*, 2024.

[22] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using A "siamese" time delay neural network. *Int. J. Pattern Recognit. Artif. Intell.*, 7(4):669–688, 1993.

[23] Alexander Brown. What is hate speech? part 1: The myth of hate. *Law and philosophy*, 36(4):419–468, 2017.

[24] Qingqing Cao, Aruna Balasubramanian, and Niranjan Balasubramanian. Towards accurate and reliable energy measurement of NLP models. In Nafise Sadat Moosavi, Angela Fan, Vered Shwartz, Goran Glavas, Shafiq R. Joty, Alex Wang, and Thomas Wolf, editors, *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing, SustaiNLP@EMNLP 2020, Online, November 20, 2020*, pages 141–148. Association for Computational Linguistics, 2020.

[25] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue

Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 321–332. Association for Computational Linguistics, 2022.

[26] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, A. J. Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali-x: On scaling up a multilingual vision and language model. *CoRR*, abs/2305.18565, 2023. URL: https://doi.org/10.48550/arXiv.2305.18565, doi:10.48550/ARXIV.2305.18565.

[27] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020.

[28] Luis Chiruzzo, Salud María Jiménez-Zafra, and Francisco Rangel. Overview of iberlef 2024: Natural language processing challenges for spanish and other iberian languages. In Salud María Jiménez-Zafra, Luis Chiruzzo, Francisco Rangel, Fazlourrahman Balouchzahi, Ulisses Brisolara Corrêa, Alba Bonet-Jover, Helena Gómez-Adorno, José Ángel González Barba, Delia Irazú Hernández Farías, Arturo Montejo-Ráez, Pablo Moral, Carlos Rodríguez Abellán, María Estrella Vallecillo Rodríguez, Mariona Taulé, and Rafael Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with the Conference of the Spanish Society for Natural Language Processing, Valladolid, Spain*, volume 3756 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

[29] Stefano Cirillo, Domenico Desiato, Giuseppe Polese, Giandomenico Solimando, Vijayan Sugumaran, and Shanmugam Sundaramurthy. Exploring the ability of emerging large language models to detect cyberbullying in social posts through new prompt-

based classification approaches. *Information Processing & Management*, 62(3):Article 104043, 2025.

[30] Charic Farinango Cuervo and Natalie Parde. Exploring contrastive learning for multimodal detection of misogynistic memes. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation, Seattle, Washington, United States*, pages 785–792. Association for Computational Linguistics, 2022.

[31] Padraig Cunningham and Sarah Delany. k-nearest neighbour classifiers. *Multiple Classifier Systems*, 54, 04 2007.

[32] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Trans. Assoc. Comput. Linguistics*, 10:92–110, 2022.

[33] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press, 2017.

[34] Pedro O. S. Vaz de Melo, Wei Jeng, and Cody Buntain, editors. *Feels Bad Man: Dissecting Automated Hateful Meme Detection Through the Lens of Facebook's Challenge*, 2022. doi:10.36190/2022.65.

[35] Suman Deep, Gurleen Kaur, Sahil Dhir, Keyur Rajyaguru, and Dimple Gajra. Green in ai: Dynamic model scaling energy- efficient model training frameworks. *IJPAR JOURNAL*, 12:13, 01 2025.

[36] Somaiyeh Dehghan, Mehmet Umut Sen, and Berrin Yanikoglu. Dealing with annotator disagreement in hate speech classification. *CoRR*, abs/2502.08266, 2025. URL: https://doi.org/10.48550/arXiv.2502.08266, arXiv:2502.08266.

[37] Tanvi Deshpande and Nitya Mani. An Interpretable Approach to Hateful Meme Detection. In *Proceedings of the International Conference on Multimodal Interaction*, ICMI '21, page 723–727, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3462244.3479949.

[38] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In Alice Oh, Tristan Naumann, Amir Globerson,

Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, New Orleans, LA, USA, December 10-16*, 2023.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[40] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[41] Ashwin Geet D'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. Label propagation-based semi-supervised learning for hate speech classification. In Anna Rogers, João Sedoc, and Anna Rumshisky, editors, *Proceedings of the First Workshop on Insights from Negative Results in NLP, Insights 2020, Online, November 19, 2020*, pages 54–59. Association for Computational Linguistics, 2020.

[42] Maeve Duggan. Online Harassment. *Pew Research Center*, 2017.

[43] Amani ElBarazi. How social media affects people's ideas on sexist behaviours and gender-based violence. 09 2023. doi:10.19080/GJIDD.2023.12.555838.

[44] European Union Agency for Fundamental Rights. Online content moderation – current challenges in detecting hate speech, 2023.

[45] Haoyi Fan, Ruidong Wang, Xunhua Huang, Fengbin Zhang, Zuoyong Li, and Shimei Su. Deep joint adversarial learning for anomaly detection on attribute networks. *Information Sciences*, 654, 2024.

[46] Jaouhar Fattahi, Feriel Sghaier, Mohamed Mejri, Ridha Ghayoula, Sahbi Bahroun, and Marwa Ziadia. Sexism discovery using cnn, word embeddings, NLP and data augmentation. In *10th International Conference on Control, Decision and Information*

Technologies, CoDIT 2024, Vallette, Malta, July 1-4, 2024, pages 1685–1690. IEEE, 2024.

[47] Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231, 2019.

[48] 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 Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 533–549. Association for Computational Linguistics, 2022.

[49] Lara Fontanella, Berta Chulvi, Elisa Ignazzi, Annalina Sarra, and Alice Tontodimamma. How do we study misogyny in the digital age? a systematic literature review using a computational linguistic approach. *Humanities and Social Sciences Communications*, 11(1):478, 2024.

[50] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30, 2018. doi:10.1145/3232676.

[51] Jesse Fox, Carlos Cruz, and Ji Young Lee. Perpetuating Online Sexism Offline: Anonymity, Interactivity, and the Effects of Sexist Hashtags on Social Media. *Computers in Human Behavior*, 52:436–442, 2015.

[52] Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28, 2024.

[53] Sushant Gautam. Bridging multimedia modalities: Enhanced multimodal AI understanding and intelligent agents. In Elisabeth André, Mohamed Chetouani, Dominique Vaufreydaz, Gale M. Lucas, Tanja Schultz, Louis-Philippe Morency, and Alessandro Vinciarelli, editors, *Proceedings of the 25th International Conference on Multimodal Interaction, ICMI 2023, Paris, France, October 9-13, 2023*, pages 695–699. ACM, 2023.

[54] Amira Ghenai, Zeinab Noorian, Hadiseh Moradisani, Parya Abadeh, Caroline Erentzen, and Fattane Zarrinkalam. Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users. *Information Processing & Management*, 62(3):Article 104079, 2025.

[55] Biagio Grasso, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. KERMIT: knowledge-empowered model in harmful meme detection. *Inf. Fusion*, 106:102269, 2024.

[56] Patricia-Carla Grigor, Bojan Evkoski, and Petra Kralj Novak. Multilingual hate speech modeling by leveraging inter-annotator disagreement. In *Slovenian KDD Conference. October 7th 2024, Ljubljana, Slovenia*, 2024.

[57] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA*, pages 770–778. IEEE Computer Society, 2016.

[58] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[59] Mika Hietanen and Johan Eddebo. Towards a definition of hate speech—with a focus on online contexts. *Journal of Communication Inquiry*, 47(4):440–458, 2023.

[60] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 2019.

[61] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29*, 2022.

[62] Zhijun Hu, Yong Xu, Jie Wen, Xian Jing Cheng, Zaijun Zhang, Lilei Sun, and Yaowei Wang. Global-supervised contrastive loss and view-aware-based post-processing for vehicle re-identification. *CoRR*, abs/2204.07943, 2022.

[63] Jianzhao Huang, Hongzhan Lin, Ziyan Liu, Ziyang Luo, Guang Chen, and Jing Ma. Towards low-resource harmful meme detection with LMM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2269–2293. Association for Computational Linguistics, 2024.

[64] Kai Huang, Hanyun Yin, Heng Huang, and Wei Gao. Towards green AI in fine-tuning large language models via adaptive backpropagation. In *The Twelfth International Conference on Learning Representations*, 2024.

[65] Elisa Ignazzi, Annalina Sarra, and Lara Fontanella. *EXPLORING MISOGYNY THROUGH TIME: FROM ITS HISTORICAL ORIGINS TO MODERN COMPLEXITIES*, pages 195–214. 12 2023.

[66] Michael P. Johnson and Kathleen J. Ferraro. Research on domestic violence in the 1990s: Making distinctions. *Journal of Marriage and the Family*, 62(2):948–963, 2000.

[67] Julia Kansok-Dusche, Cindy Ballaschk, Norman Krause, Anke Zeißig, Lisanne Seemann-Herz, Sebastian Wachs, and Ludwig Bilz. A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, Violence, & Abuse*, 24(4):2598–2615, 2023.

[68] Prashant Kapil and Asif Ekbal. A deep neural network based multi-task learning approach to hate speech detection. *Knowl. Based Syst.*, 210:106458, 2020.

[69] S. Karishma and V. Akila. A comparative analysis of multimodal misogyny memes using deep learning with semi supervised learning algorithms. In *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pages 1791–1796, 2025.

[70] Shakir Khan, Mohd Fazil, Vineet Kumar Sejwal, Mohammed Ali Alshara, Reemiah Muneer Alotaibi, Ashraf Kamal, and Abdul Rauf Baig. Bichat: Bilstm with deep CNN and hierarchical attention for hate speech detection. *J. King Saud Univ. Comput. Inf. Sci.*, 34(7):4335–4344, 2022.

[71] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan,

and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[72] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[73] Kyeong-Hwan Kim and Chang-Sung Jeong. F-albert: A distilled model from a two-time distillation system for reduced computational complexity in albert model. *Applied Sciences*, 13(17), 2023.

[74] Gokul Karthik Kumar and Karthik Nandakumar. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features. *CoRR*, abs/2210.05916, 2022. URL: https://doi.org/10.48550/arXiv.2210.05916, arXiv:2210.05916, doi:10.48550/ARXIV.2210.05916.

[75] Ayşe I. Kural and Monika Kovács. Attachment security schemas to attenuate the appeal of benevolent sexism: The effect of the need to belong and relationship security. *Acta Psychologica*, 229:103671, 2022.

[76] Gretel Liz De la Peña Sarracén and Paolo Rosso. Systematic keyword and bias analyses in hate speech detection. *Information Processing & Management*, 60(5):Article 103433, 2023.

[77] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[78] Jane Kathrine Larsen. Sexism and misogyny in american hip-hop culture, 2006. URL: https://www.duo.uio.no/handle/10852/25447.

[79] Phuong Le-Hong. Diacritics generation and application in hate speech detection on vietnamese social networks. *Knowl. Based Syst.*, 233:107504, 2021.

[80] Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. Semeval-2023 task 11: Learning with disagreements (lewidi). In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada*, pages 2304–2318. Association for Computational Linguistics, 2023.

[81] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900, 2022.

[82] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705, 2021.

[83] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. URL: http://arxiv.org/abs/1908.03557, arXiv:1908.03557.

[84] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics (ACL)*, pages 5265–5275, 2020.

[85] Xianshuai Li, Zhi Liu, and Sannyuya Liu. Triple contrastive learning representation boosting for supervised multiclass tasks. *Information Processing & Management*, 62(3):Article 104011, 2025.

[86] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceed-*

ings, Part XXX, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020.

[87] Baohao Liao, Yan Meng, and Christof Monz. Parameter-efficient fine-tuning without introducing new latency. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4242–4260. Association for Computational Linguistics, 2023.

[88] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. A multimodal framework for the detection of hateful memes, 2020. arXiv:2012.12871.

[89] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL: http://arxiv.org/abs/1907.11692, arXiv:1907.11692.

[90] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982.

[91] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019.

[92] Junyu Lu, Hongfei Lin, Xiaokun Zhang, Zhaoqing Li, Tongyue Zhang, Linlin Zong, Fenglong Ma, and Bo Xu. Hate speech detection via dual contrastive learning. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2787–2795, 2023.

[93] Junyu Lu, Bo Xu, Xiaokun Zhang, Haohao Zhu, Kaichun Wang, Liang Yang, and Hongfei Lin. Is having rationales enough? rethinking knowledge enhancement for multimodal hateful meme detection. In Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne, editors, *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 559–569. ACM, 2025.

[94] Alexandra Luccioni, Alexandre Lacoste, and Victor Schmidt. Estimating carbon emissions of artificial intelligence [opinion]. *IEEE Technol. Soc. Mag.*, 39(2):48–51, 2020.

[95] Alexandra Sasha Luccioni and Alex Hernández-García. Counting carbon: A survey of factors influencing the emissions of machine learning. *CoRR*, abs/2302.08476, 2023. doi:10.48550/ARXIV.2302.08476.

[96] Florian Ludwig, Klara Dolos, Ana Alves-Pinto, and Torsten Zesch. Unraveling the dynamics of semi-supervised hate speech detection: The impact of unlabeled data characteristics and pseudo-labeling strategies. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 1974–1986. Association for Computational Linguistics, 2024.

[97] Jitendra Singh Malik, Guansong Pang, and Anton van den Hengel. Deep learning for hate speech detection: A comparative study. *CoRR*, abs/2202.09517, 2022. URL: https://arxiv.org/abs/2202.09517, arXiv:2202.09517.

[98] Verma Ishita Manoj. Lightweight machine learning models with python for green ai. *International Journal of Multidisciplinary Research in Science, Engineering, Technology and Management*, 11(6), 2024.

[99] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders, editors, *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 1485–1488. ACM, 2010. doi:10.1145/1873951.1874254.

[100] Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. Predicting hate intensity of twitter conversation threads. *Knowl. Based Syst.*, 275:110644, 2023.

[101] Khouloud Mnassri, Reza Farahbakhsh, and Noël Crespi. Multilingual hate speech detection: A semi-supervised generative adversarial approach. *Entropy*, 26(4):344, 2024.

[102] Yufei Mu, Jin Yang, Tianrui Li, Siyu Li, and Weiheng Liang. HA-GCEN: hyperedge-abundant graph convolutional enhanced network for hate speech detection. *Knowl. Based Syst.*, 300:112166, 2024.

[103] Raymond T Mutanga, Nalindren Naicker, and Oludayo O Olugbara. Hate speech detection in twitter using transformer methods. *International Journal of Advanced Computer Science and Applications*, 11(9), 2020.

[104] Chikashi Nobata, Joel R. Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In Jacqueline Bourdeau, Jim

Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 145–153. ACM, 2016.

[105] Petra Kralj Novak, Teresa Scantamburlo, Andraz Pelicon, Matteo Cinelli, Igor Mozetic, and Fabiana Zollo. Handling disagreement in hate speech modelling. In Davide Ciucci, Inés Couso, Jesús Medina, Dominik Slezak, Davide Petturiti, Bernadette Bouchon-Meunier, and Ronald R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems - 19th International Conference, IPMU 2022, Milan, Italy, July 11-15, 2022, Proceedings, Part II*, volume 1602 of *Communications in Computer and Information Science*, pages 681–695. Springer, 2022.

[106] Luca Oneto, Sandro Ridella, and Davide Anguita. Informed machine learning: Excess risk and generalization. *Neurocomputing*, 646:130521, 2025.

[107] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):Article 102360, 2020.

[108] Ronghao Pan, José Antonio García Díaz, Tomás Bernal Beltrán, and Rafael Valencia-García. Umuteam at exist 2024: multi-modal identification and categorization of sexism by feature integration. *Working Notes of CLEF*, 2024.

[109] Shashank Pareek, Ahmed Saleh Al-Samalek, Ahmed Alkhayyat, Sandeep Singh, Amrita Singh, and Shivakrishna Dasi. Efficient vision transformers for edge devices: Pruning and quantization approaches. In *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pages 1465–1471, 2024.

[110] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019.

[111] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the

world. *CoRR*, abs/2306.14824, 2023. URL: https://doi.org/10.48550/arXiv.2306.14824, arXiv:2306.14824, doi:10.48550/ARXIV.2306.14824.

[112] Laura Plaza, Jorge Carrillo-de-Albornoz, Enrique Amigó, Julio Gonzalo, Roser Morante, Paolo Rosso, Damiano Spina, Berta Chulvi, Alba Maeso, and Víctor Ruiz. EXIST 2024: sexism identification in social networks and memes. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V*, volume 14612 of *Lecture Notes in Computer Science*, pages 498–504. Springer, 2024.

[113] Laura Plaza, Jorge Carrillo-de-Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. Overview of EXIST 2023: sexism identification in social networks. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, volume 13982 of *Lecture Notes in Computer Science*, pages 593–599. Springer, 2023.

[114] Laura Plaza, Jorge Carrillo de Albornoz, Víctor Ruiz, Alba Maeso, Berta Chulvi, Paolo Rosso, Enrique Amigó, Julio Gonzalo, Roser Morante, and Damiano Spina. Overview of EXIST 2024 - learning with disagreement for sexism identification and characterization in tweets and memes (extended overview). In Guglielmo Faggioli, Nicola Ferro, Petra Galuscáková, and Alba García Seco de Herrera, editors, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 908–941, 2024.

[115] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4439–4455. Association for Computational Linguistics, 2021.

[116] Cendra Devayana Putra and Hei-Chia Wang. Semi-meta-supervised hate speech detection. *Knowl. Based Syst.*, 287:111386, 2024.

[117] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[118] Abir Rahali, Moulay A. Akhloufi, Anne-Marie Therien-Daniel, and Éloi Brassard-Gourdeau. Automatic misogyny detection in social media platforms using attention-based bidirectional-lstm. In *2021 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2021, Melbourne, Australia, October 17-20, 2021*, pages 2706–2711. IEEE, 2021.

[119] Anchal Rawat, Santosh Kumar, and Surender Singh Samant. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*, 16, 2024. doi:10.1002/wics.1648.

[120] Mohammad Zia Ur Rehman, Somya Mehta, Kuldeep Singh, Kunal Kaushik, and Nagendra Kumar. User-aware multilingual abusive content detection in social media. *Information Processing & Management*, 60(5):Article 103450, 2023.

[121] Mohammad Zia Ur Rehman, Aditya Shah, and Nagendra Kumar. An adaptive supervised contrastive learning framework for implicit sexism detection in digital social networks. *CoRR*, abs/2507.05271, 2025. URL: https://doi.org/10.48550/arXiv.2507.05271, arXiv:2507.05271, doi:10.48550/ARXIV.2507.05271.

[122] Mohammad Zia Ur Rehman, Sufyaan Zahoor, Areeb Manzoor, Musharaf Maqbool, and Nagendra Kumar. A context-aware attention and graph neural network-based multimodal framework for misogyny detection. *Inf. Process. Manag.*, 62(1):103895, 2025.

[123] Gang Ren, Li Jiang, Tingting Huang, Ying Yang, and Taeho Hong. Temporal-spatial hierarchical contrastive learning for misinformation detection: A public-behavior perspective. *Information Processing & Management*, 62(4):Article 104108, 2025.

[124] Michael Ridenhour, Arunkumar Bagavathi, Elaheh Raisi, and Siddharth Krishnan. Detecting online hate speech: Approaches using weak supervision and network embedding models. In Robert Thomson, Halil Bisgin, Christopher L. Dancy, Ayaz Hyder, and Muhammad Hussain, editors, *Social, Cultural, and Behavioral Modeling -*

13th International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18-21, 2020, Proceedings, volume 12268 of *Lecture Notes in Computer Science*, pages 202–212. Springer, 2020.

[125] Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(5):Article 103474, 2023.

[126] Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini1. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia, 2024.

[127] Giulia Rizzi, Paolo Rosso, and Elisabetta Fersini. From explanation to detection: Multimodal insights into disagreement in misogynous memes. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2024.

[128] Giulia Rizzi, Paolo Rosso, and Elisabetta Fersini. Is a bunch of words enough to detect disagreement in hateful content? In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025 - Workshops, Abu Dhabi, UAE, January 19-24, 2025*, pages 1–11. Association for Computational Linguistics, 2025.

[129] Francisco Rodríguez-Sánchez, Jorge Carrillo-de-Albornoz, and Laura Plaza. Leveraging unsupervised task adaptation and semi-supervised learning with semantic-enriched representations for online sexism detection. *Expert Syst. J. Knowl. Eng.*, 42(2), 2025.

[130] Yisi Sang and Jeffrey M. Stanton. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In Malte Smits, editor, *Information for a Better World: Shaping the Global Future - 17th International Conference, iConference 2022, Virtual Event, Proceedings, Part I*, volume 13192 of *Lecture Notes in Computer Science*, pages 425–444. Springer, 2022.

[131] Ranjan Sapkota and Manoj Karkee. Object detection with multimodal large vision-language models: An in-depth review. *Inf. Fusion*, 126:103575, 2026.

[132] Francesco Scala, Sergio Flesca, and Luigi Pontieri. Play it straight: An intelligent data pruning technique for green-ai. In Dino Pedreschi, Anna Monreale, Riccardo Guidotti, Roberto Pellungrini, and Francesca Naretto, editors, *Discovery Science - 27th International Conference, DS 2024, Pisa, Italy, October 14-16, 2024, Proceedings, Part I*, volume 15243 of *Lecture Notes in Computer Science*, pages 69–85. Springer, 2024.

[133] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[134] Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

[135] Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. AOMD: an analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58(5):Article 102664, 2021.

[136] Arushi Sharma, Anubha Kabra, and Minni Jain. Ceasing hate with *MoH*: Hate speech detection in hindi-english code-switched language. *Information Processing & Management*, 59(1):Article 102760, 2022.

[137] Tony Kim Smith, H Ruda Nie, Johanne R Trippas, and Damiano Spina. Rmit-ir at exist lab at clef 2024. *Working Notes of CLEF*, 2024.

[138] Rui Song, Fausto Giunchiglia, Yingji Li, Jian Li, Jingwen Wang, and Hao Xu. KALD: A knowledge augmented multi-contrastive learning model for low resource abusive language detection. *Knowl. Based Syst.*, 321:113619, 2025.

[139] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics, 2019.

[140] Xuanyu Su, Yansong Li, Diana Inkpen, and Nathalie Japkowicz. A context-aware contrastive learning framework for hateful meme detection and segmentation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5201–5215. Association for Computational Linguistics, 2025.

[141] Janet K. Swim, Lauri L. Hyers, Laurie L. Cohen, Davita C. Fitzgerald, and Wayne H. Bylsma. African american college students' experiences with everyday racism: Characteristics of and responses to these incidents. *Journal of Black Psychology*, 29(1):38–67, 2003. doi:10.1177/0095798402239228.

[142] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019.

[143] The Council of Europe. Grevio general recommendation no. 1 on the digital dimension of violence against women, 2021.

[144] Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. Semeval-2021 task 12: Learning with disagreements. In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurélie Herbelot, and Xiaodan Zhu, editors, *Proceedings of the 15th International Workshop on Semantic Evaluation, Virtual Event, Bangkok, Thailand*, pages 338–347. Association for Computational Linguistics, 2021.

[145] United Nations. What is hate speech? https://www.un.org/en/hate-speech/. Accessed: 2026-01-01.

[146] Lorenzo Vaiani, Luca Cagliero, Paolo Garza, and Jason Ravagli. Cross-modal consistency types in multimodal social data. *Knowl. Based Syst.*, 322:113705, 2025.

[147] Advaitha Vetagiri, Partha Pakray, and Amitava Das. A deep dive into automated sexism detection using fine-tuned deep learning and large language models. *Engineering Applications of Artificial Intelligence*, 145:Article 110167, 2025.

[148] Emily A. Vogels. The state of online harassment, 2021.

[149] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 577–584. Morgan Kaufmann, 2001.

[150] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *Trans. Mach. Learn. Res.*, 2022, 2022.

[151] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics, 2016.

[152] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020.

[153] Samantha Pinson Wrisley. Feminist theory and the problem of misogyny. *Feminist Theory*, 24(2):188–207, 2023.

[154] Fan Wu, Bin Gao, Xiaoou Pan, Linlin Li, Yujiao Ma, Shutian Liu, and Zhengjun Liu. Fuser: An enhanced multimodal fusion framework with congruent reinforced perceptron for hateful memes detection. *Information Processing & Management*, 61(4):Article 103772, 2024.

[155] Chuanpeng Yang, Fuqing Zhu, Yaxin Liu, Jizhong Han, and Songlin Hu. Uncertainty-aware cross-modal alignment for hate speech detection. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16973–16983. ELRA and ICCL, 2024.

[156] Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. HARE: explainable hate speech detection with step-by-step reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computa-*

tional Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, pages 5490–5505. Association for Computational Linguistics, 2023.

[157] Lanqin Yuan, Tianyu Wang, Gabriela Ferraro, Hanna Suominen, and Marian-Andrei Rizoiu. Transfer learning for hate speech detection in social media. *Journal of Computational Social Science*, 6(2):1081–1101, 2023.

[158] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE, 2021.

[159] Gang Zhou, Haizhou Wang, Di Jin, Wenxian Wang, Shuyu Jiang, Rui Tang, and Xingshu Chen. A toxic euphemism detection framework for online social network based on semantic contrastive learning and dual channel knowledge augmentation. *Information Processing & Management*, 62(5):Article 104143, 2025.

[160] Reinhard Zimmermann. *The Law of Obligations: Roman Foundations of the Civilian Tradition*. Oxford University Press, 08 1996.

Tesi di dottorato realizzata nell'ambito del progetto ISCS finanziato dal PNRR Missione 4 Componente 2 Investimento 1.4, finanziato dall'Unione Europea – NextGenerationEU – CUP H43C22000520001