VRIJE UNIVERSITEIT AMSTERDAM

Master Thesis

Evaluating the Impact of Linguistic Features in Harmful Meme Detection: A Systematic Ablation Study

Shenglin Li

Supervisor Ilia Markov
2^{nd} reader Piek Vossen

a thesis submitted in fulfillment of the requirements for the degree of

MA Linguistics
(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

Date June 27, 2025
Student number 2838318
Word count 14126

Abstract

Memes have become a popular and influential form of online communication. Although often humorous, they can also serve as vehicles for harmful content, including misogynistic and sexist messages. The multimodal nature and implicit expression of memes make them particularly challenging for automatic detection systems such as content moderation on social platforms.

To address these challenges and better understand the factors that influence detection model performance in identifying harmful content, this thesis investigates the contribution of linguistic features to harmful meme detection, focusing on misogynistic and sexist content across two datasets: MAMI and EXIST2024. The memes were represented in terms of meme text and image caption (generated with the BLIP-2 model), after which a systematic ablation study was conducted on four key feature categories: sentiment markers, emotion-based words, function words, and hate speech lexicon terms for both SVM and BERT models. Binary classification experiments revealed that these features had limited influence on coarse-grained performance. To explore which word categories the models rely on, I conducted part-of-speech (POS) ablation experiments showing that content words such as nouns and proper nouns are important for misogynistic and sexist meme detection, likely because of their high frequency in the input text.

In multi-label classification with SVM, negative sentiment emerged as the most impactful coarse-grained feature category. Fine-grained ablation further revealed that different negative emotions served distinct roles in identifying harm subtypes: sadness-related words had the strongest impact on detecting shaming and violence in misogynistic detection, whereas anger-related words were especially predictive of sexual-violence and misogyny-non-sexual-violence in sexist detection. Function words, specifically pronouns, were particularly important for identifying fine-grained objectification, probably because they help the model determine who is the target of the misogynous/sexist content and between in-group and out-group references.

These findings highlight the importance of specific linguistic features in harmful meme detection and reveal the complex interdependencies between feature categories and harm subtypes. Through detailed ablation analysis, this study also emphasizes challenges in multimodal modeling and proposes directions for future research.

Keywords:

Harmful meme detection; Misogyny and sexism; Ablation study; Multimodal classification; Linguistic features; Traditional machine learning models; Transformer-based models

Declaration of Authorship

I, Shenglin Li, declare that this thesis, titled *Evaluating the Impact of Linguistic Features in Harmful Meme Detection: A Systematic Ablation Study* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master's degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: <28-06-2025>

Signed: 

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Ilia Markov, for his continuous support, insightful feedback, and invaluable guidance throughout the process of this thesis. His expertise in computational linguistics and thoughtful mentorship have been crucial to the development of this research.

I am also thankful to the faculty and staff of the CLTL Lab at VU Amsterdam for providing a stimulating academic environment and for their support throughout my Master's studies.

My heartfelt thanks go to my peers Cheryl Chen, Susana Chen, and Melina Paxinou, whose encouragement, thoughtful discussions, and companionship made this journey all the more meaningful in both this thesis project and the master's program experienced together.

Finally, I am deeply grateful to my family and best friends, whose support and belief in me gave me strength during the most challenging times. This work would not have been possible without your presence by my side.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Social media like Twitter, Instagram, and TikTok has changed how people interact and exchange ideas online. Among the various forms of digital content that have emerged, memes stand out as particularly influential since they have moved beyond mere entertainment to become a powerful and widely used medium for expressing ideas (Shifman, 2014; Wiggins, 2019). Memes, while commonly used to share jokes and shape individual and group identity (Lestari et al., 2024), have also been used to disseminate harmful content, in particular against marginalized communities. A particularly harmful form of this behavior comes in the form of misogynous/sexist memes, which reinforce or even deepen negative stereotypes and promote toxic spaces for women in the digital world (Fersini et al., 2022; Kiela et al., 2020). Automated detection of such harmful content poses serious technical and theoretical challenges beyond the realm of the current content moderation approaches based on textual features, as memes often rely on implicit meanings not explicitly conveyed through surface-level text and images (Lin et al., 2024). Moreover, many memes derive their harmful impact from the complex interaction between textual and visual components, creating meaning that neither modality conveys on its own.

The multimodal nature of memes, which makes the meaning created within of the complex interaction between text and visual content (Pramanick et al., 2021), creates additional challenges for harmful meme detection. Unlike plain textual harmful content, memes leverage multimodal relationship to convey harmful messages through subtle linguistic cues and implicit messaging that often evade conventional detection systems (Fersini et al., 2022). Furthermore, the subjective nature of humor and offense introduces additional layers of complexity, as identical content may be interpreted differently across diverse contexts and audiences (Williams et al., 2020).

Understanding the specific linguistic strategies underlying harmful memes is crucial not only for improving detection accuracy but also for developing interpretable models that can explain their classification decisions to human moderators. Despite substantial advances in multimodal approaches to harmful meme detection, a critical gap remains in our understanding of which specific linguistic features contribute most indicative to model performance. While the effectiveness of natural language processing (NLP) models largely depends on their capacity to identify and leverage relevant linguistic patterns, the relative importance of different feature categories such as sentiment markers, emotion words, function words, and hate speech lexicon terms remains systematically underexplored in the context of misogynous and sexist meme detection.

This research addresses this knowledge gap through systematic ablation studies de-

signed to evaluate the differential impact of linguistic feature categories on harmful meme detection performance. By focusing specifically on binary and multi-label classification experiments on misogynous and sexist content, this investigation aims to enhance model explainability, illuminate the linguistic mechanisms underlying harmful memes, and inform the development of more efficient and targeted content moderation systems. Additionally, the study examines potential differences in linguistic patterns between misogynous and sexist content classifications. This analysis has the potential to uncover distinctive linguistic features that distinguish these related but conceptually distinct types of harmful content. Through systematic ablation studies across multiple feature categories and two datasets, this research provides empirical evidence for differential feature contributions.

1.1 Research Questions and Objectives

To systematically investigate these linguistic mechanisms, this research is guided by the following research questions:

Primary Research Question

- Which features are the most indicative for harmful meme detection across different datasets, experimental setups (binary and multi-label) and models (SVM and BERT)?

Research Subquestions

- How do positive and negative sentiment markers contribute to harmful meme classification performance?
- What is the role of emotion-based features (e.g., anger, fear, joy) in harmful meme detection?
- To what extent do function words influence model performance in harmful meme detection?
- Which specific categories of hate speech lexicon are most predictive of harmful content?

1.2 Thesis Organization

This thesis is structured as follows:

- **Chapter 2** reviews existing literature on harmful content detection, multimodal approaches, and feature importance analysis.
- **Chapter 3** details the datasets, ablation study methodology, experimental design and evaluation framework.
- **Chapter 4** presents comprehensive results from systematic binary and multi-label ablation studies across different feature categories for both datasets and models, and provides detailed error analysis for the most indicative feature categories.

- **Chapter 5** summarizes the most important findings, addresses study limitations, and provides directions for future research.

The findings of this research have important implications for both academic understanding of harmful content linguistics and practical applications in content moderation, contributing to more effective and explainable systems for detecting harmful memes in digital environments.

Chapter 2

Related Work

This chapter provides an overview of the key developments in harmful content detection. It begins by tracing how harmful content detection has evolved from early keyword-based methods to modern multimodal systems. Next, it reviews two shared tasks MAMI, which target misogynous memes, and EXIST2024, which focuses on sexist content. This chapter also discusses four different types of linguistic features used in past research: sentiment and emotion-based words, function words, and hate speech lexicons. In addition, it covers recent efforts in multimodal analysis, especially those that use image captioning to link textual and visual components. Finally, the chapter highlights the role of ablation studies in understanding feature importance and identifies research gaps that motivate this thesis.

2.1 Evolution of Harmful Content Detection

The detection of harmful content has evolved from early keyword-based approaches to sophisticated multimodal systems. Initial research efforts focused primarily on lexical features, leveraging traditional machine learning techniques such as Support Vector Machines (SVM), Naive Bayes (NB) and Logistic Regression (LR) to identify hate speech or offensive content based on the presence of known hateful and abusive terms (Davidson et al., 2017; Schmidt and Wiegand, 2017). These systems often utilize bag-of-words (BoW), n-gram, or term frequency–inverse document frequency (TF-IDF) representations to encode textual input. While such representations offer reasonable performance for detection, they are limited in capturing deeper contextual nuances inherent in more implicit forms of harmful content.

The introduction of deep learning brings a substantial improvement in harmful content detection. Badjatiya et al. (2017) demonstrated that neural models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) networks, can outperform traditional approaches by capturing more complex linguistic structures and contextual dependencies. However, these models remain confined to textual inputs and face limitations when addressing the multimodal nature of contemporary online content.

Recognition of the multimodal character of harmful content has led to the development of integrated approaches that incorporate both textual and visual information. (Kiela et al., 2020) introduced the Hateful Memes Challenge, which emphasized the need to jointly analyze text and image to detect memes that communicate harmful messages through their combined modalities. Their findings confirmed that multi-

modal models did better than unimodal baselines, thereby establishing the importance of multimodal fusion in harmful content detection.

2.2 Shared Tasks on Misogyny and Sexism Detection

In recent years, shared tasks have played a pivotal role in advancing research on misogynous and sexist content detection. The SemEval-2022 Task 5, known as the Multimedia Automatic Misogyny Identification (MAMI) challenge, provides a dedicated benchmark for identifying misogynous memes by combining image and text modalities (Fersini et al., 2022). This task consists of two subtasks: a binary classification of misogynous versus non-misogynous memes, and a fine-grained multi-label classification across four categories of misogyny: stereotyping, objectification, shaming, and violence. The top-performing system designed by SRCB (Zhang and Wang, 2022) ranked first in subtask A with macro-F1 scores of 0.834. It also achieved the best result (0.731) on Subask B along with TIB-VA' (Hakimov et al., 2022) and PAIC' (Zhi et al., 2022) models. These results highlight the importance of effective multimodal integration for handling fine-grained harmful meme classification.

The EXIST shared task series has similarly contributed to research in sexist content detection. The 2024 edition expanded the scope to include multimodal content, particularly memes, in addition to traditional tweet-based data (Plaza-Del-Arco et al., 2020). The task involvs both binary (Task 4) and multi-label classification (Task 3) in English and Spanish, using a Learning with Disagreement (LeWiDi) paradigm to reflect annotator diversity. According to Plaza et al. (2024), the top-performing team in the binary task was Victor-UNED_1, which achieved the highest ICM-SoftNorm score of 0.4530. For the multi-label task, NYCU-NLP_1 ranked first with an ICM-Soft score of –1.1762 and a normalized score of 0.4379. These results highlight both the effectiveness of transformer-based models and the challenges of achieving robust performance across different modalities and languages.

2.3 Linguistic Features in Harmful Content Detection

Linguistic features play a vital role in detecting harmful content, particularly in contexts where explicit signals are subtle or absent. Among these, sentiment indicators and emotion-based features have gained substantial attention.

Sentiment polarity analysis has been widely applied in harmful content detection, with researchers leveraging both positive and negative sentiment scores to identify potentially harmful messages. Njagi et al. (2015) demonstrated that sentiment polarity features, particularly strongly negative sentiment expressions, can effectively distinguish hate speech from regular content by analyzing the subjectivity and polarity of text. Their lexicon-based approach showed that combining sentiment polarity scores with thematic features significantly improved detection performance. Similarly, Naznin et al. (2024) found that hierarchical sentiment analysis frameworks considerably enhanced hate speech detection by incorporating sentiment classification as a preprocessing step to reduce false positives.

Complementing sentiment analysis, emotion-based features have also proven valuable. The NRC Emotion Lexicon, developed by Mohammad and Turney (2013), provides associations between approximately 14,000 English words and eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. This resource

has been widely used in sentiment and emotion analysis and has proven effective in identifying the emotional framing of harmful messages. Plaza-Del-Arco et al. (2020) demonstrated that incorporating emotion features improved misogyny detection, especially when combined with other linguistic cues. However, the specific contribution of individual emotion categories to harmful content detection remains underexplored, both for binary classification (harmful vs. non-harmful) and for distinguishing between different subtypes of harmful content.

Another important category is function words, which include closed-class words such as determiners, prepositions, conjunctions, pronouns, and auxiliary verbs. These words are often overlooked as they have very little meaning. However, stylometric studies suggest function words can reveal consistent patterns of language use. Markov et al. (2021) found that function word distributions were predictive of hate speech across domains and languages. Moreover, they can be used to express divisive opinions between the in-group and the out-group community (Alorainy et al., 2019). Drawing on the psychological framework developed by Pennebaker (2011), researchers have demonstrated that function words can capture stylistic and cognitive aspects of harmful communication. Nevertheless, the specific role of function words in multimodal settings, particularly in distinguishing between fine-grained harmful categories, remains an open research question that warrants systematic investigation.

Hate speech lexicons offer another structured approach to feature extraction. HurtLex (Bassignana et al., 2018), a multilingual lexicon with 17 offensive language categories, enables fine-grained categorization beyond simple profanity filtering. Chiril et al. (2022) demonstrated that category-specific lexical features derived from HurtLex enhanced model performance over general-purpose offensive term lists. While hate speech lexicons are particularly effective at capturing explicit harmful content rather than implicit harmful messaging, these features warrant exploration in multimodal meme detection to assess their potential impact alongside other linguistic indicators. Despite these advances, the application of such lexicons in multimodal meme detection remains limited and demands further empirical analysis.

2.4 Multimodal Linguistic Analysis

Recent advances in multimodal analysis have enabled the exploration of textual and visual interactions in harmful memes. Pramanick et al. (2021) introduced the HarMeme dataset and showed that harmful memes often rely on the interplay between image and text, which necessitates joint modeling. Their experiments with models such as VisualBERT and ViLBERT revealed that multimodal fusion improves performance, but also exposed issues of model bias and limited interpretability.

To bridge modalities, image captioning models such as Bootstrapping Language-Image Pre-training (BLIP) (Li et al., 2022) and BLIP-2 (Li et al., 2023) have been employed to extract textual descriptions of images, facilitating linguistic analysis. This approach enables direct comparison between meme text and visual content in a unified text-based framework. Sharma et al. (2022) proposed attention-based fusion strategies to better capture cross-modal interactions. However, the nuanced relationship between linguistic features in meme text and image captions remains underexplored, particularly regarding how these features contribute to specific types of harmful intent. This gap in multimodal linguistic analysis motivates the systematic investigation of feature interactions between textual and visual modalities in the present study.

2.5 Ablation Studies and Feature Importance

Ablation studies have become a foundational methodology in NLP to evaluate the contribution of specific features or model components. In harmful content detection, Caselli et al. (2021) employed ablation techniques to assess model robustness under different input perturbations. While architectural ablation, such as layer-wise BERT analysis (Rogers et al., 2020), is common, systematic ablation of linguistic features remains scarce.

For traditional models like SVMs, features can be removed or replaced with neutral placeholders to measure their contribution. In transformer-based models, ablation can be conducted by masking tokens (e.g., with [MASK]) or omitting them entirely. Van Nooten et al. (2021) demonstrated this approach by systematically removing entire word classes to evaluate their impact on age detection models, showing how ablation studies can reveal the differential importance of linguistic features across both traditional machine learning and transformer-based approaches. Each method offers different insights into model dependencies. Despite their interpretability potential, few studies have applied controlled ablation of sentiment and emotion-based words, function words, or hate lexicon terms to multimodal content detection.

2.6 Research Gaps Addressed by This Study

Despite substantial progress, several gaps remain. First, while emotion-based features are widely used, their category-specific importance for harmful content detection has not been systematically evaluated, both for binary classification (harmful vs. non-harmful) and for fine-grained harmful content detection (e.g., distinguishing shaming from violence). Second, the role of function words in multimodal contexts remains poorly understood, particularly their interaction with visual content and their impact on both binary and fine-grained classification. Even though previous studies point in the direction of the existence of a linguistic register for hate speech messages with specific stylistic properties, the contribution of these stylometric and emotion-word features in multimodal settings requires systematic investigation. Third, although ablation studies are standard for model interpretability, few have been applied to evaluate the linguistic features central to harmful content. Fourth, comparative studies across datasets like MAMI and EXIST2024 are rare, limiting understanding of generalizability across datasets. Fifth, there is limited exploration of how linguistic patterns differ between meme text and generated image captions. Finally, while traditional machine learning approaches like SVMs and modern transformer-based models like BERT have both been applied to harmful content detection, systematic comparison of how these different models respond to linguistic feature ablation remains underexplored. It should be noted that direct comparison between these different approaches has inherent limitations - SVM largely depends on direct feature input, while BERT utilizes embedded pre-trained representations.

This thesis systematically addresses these gaps by conducting comprehensive ablation studies to examine the contribution of four key linguistic feature categories (sentiment markers, emotion-based words, function words, and hate speech lexicon terms) to misogynous/sexist detection. By examining performance across binary and multi-label classification settings on the MAMI and EXIST2024 datasets using both SVM and BERT models, the study contributes to a deeper understanding of multimodal harmful

content detection and the linguistic patterns that underpin it.

Chapter 3

Methodology

3.1 Datasets

3.1.1 MAMI Dataset

The **MAMI (Multimodal Abuse and Misogyny Identification)** dataset, introduced as part of SemEval-2022 Task 5 (Fersini et al., 2022), serves as the primary dataset for analyzing misogynistic content. The dataset contains 11,000 samples: 10,000 training samples and 1,000 test samples (shown in Table 3.1). For binary classification, the dataset maintains perfect balance with 50% *misogynous* and 50% *non-misogynous* content. For multi-label classification, memes must be categorized into one or more specific subcategories of the following four misogynistic categories (Fersini et al., 2022):

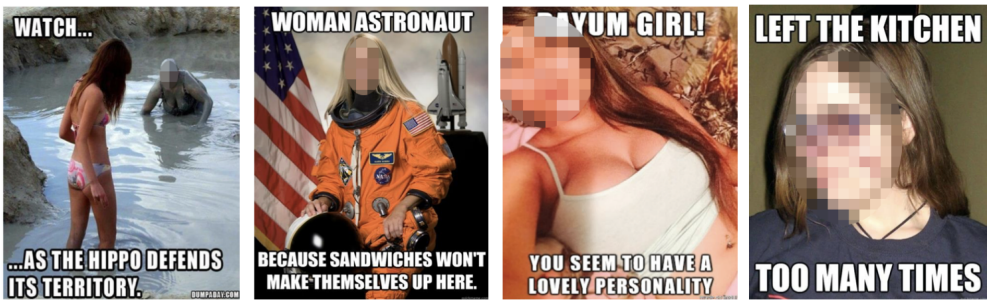
- **Shaming**: Content that criticizes women who violate expectations of behavior and appearance regarding gender typology or physical appearance, seeking to insult and offend women based on body or personality characteristics.

- **Stereotype**: Content presenting fixed, conventional ideas or characteristics assigned to women, including role stereotyping based on societal roles or gender stereotyping related to personality traits and domestic behaviors.

- **Objectification**: Content that treats women as objects, reducing them to their physical attributes rather than recognizing their dignity and personal aspects.

- **Violence**: Content that indicates physical violence and/or calls to violence against women.

Figure 3.1 presents examples of memes corresponding to each of the four misogynous subcategories, illustrating the visual and textual cues associated with each class.
Note: The memes are pixilated for privacy reasons.



(a) Shaming (b) Stereotype (c) Objectification (d) Violence

Figure 3.1: Examples of memes from the different misogynous categories.

The distribution (shown in Table 3.1) shows that *stereotyping* and *objectification* are the dominant categories, while *violence* represents the smallest category.

Label Category	**Training**			**Test**		
	Count	%	Total	Count	%	Total
Binary Negative	5,000	50.0%	10,000	500	50.0%	1,000
Binary Positive	5,000	50.0%		500	50.0%	
Shaming	1,274	12.7%		146	14.6%	
Stereotype	2,810	28.1%		350	35.0%	
Objectification	2,202	22.0%		348	34.8%	
Violence	953	9.5%		153	15.3%	

Table 3.1: MAMI Label Distribution (Binary and Fine-grained)

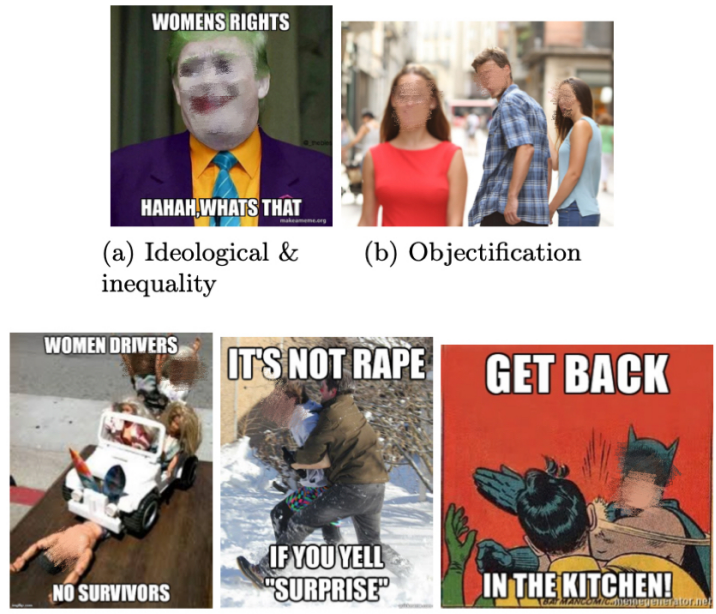
3.1.2 EXIST2024 Dataset

As shown in Table 3.2 , the **EXIST2024 (sEXism Identification in Social neTworks task at CLEF 2024)** focuses on detecting sexist content across social networks and contains approximately 1,700 samples (1,530 training and 171 test samples). The dataset exhibits a slight bias toward sexist content and includes five fine-grained multi-label categories:

- **Ideological and inequality**: Content that discredits the feminist movement, rejects the existence of inequality between men and women, or portrays men as victims of gender-based oppression.
- **Stereotyping and dominance**: Content expressing false beliefs about women, suggesting they are more suitable for certain roles or incapable of performing certain tasks, or claiming the superiority of men.
- **Objectification**: Content that presents women as objects, disregarding their dignity and personal characteristics, or describing physical traits that women must possess to conform to traditional gender roles.
- **Sexual violence**: Content that includes sexual innuendo, requests for sexual favors, sexual harassment, or references to rape or sexual assault.

- **Misogyny and non-sexual violence**: Content that conveys hatred, hostility, or incites violence toward women without explicit sexual context.

Figure 3.2 shows representative examples of memes from the EXIST2024 dataset. *Note: The memes are pixilated for privacy reasons.*



(a) Ideological & inequality (b) Objectification

(c) Stereotyping & dominance (d) Sexual violence (e) Misogyny & non-sexual violence

Figure 3.2: Examples of memes from the different sexist categories (Plaza et al., 2024).

The fine-grained categories in EXIST2024 show a more even distribution compared to MAMI, with *stereotyping-dominance* and *objectification* being most prevalent, while both *sexual-violence* and *misogyny-non-sexual-violence* categories represent approximately 10% each.

Label Category	**Training**			**Test**		
	Count	%	Total	Count	%	Total
Binary Negative	667	43.6%	1,530	76	44.4%	171
Binary Positive	863	56.4%		95	55.6%	
Ideological-inequality	369	24.1%		39	22.8%	
Stereotyping-dominance	440	28.8%		40	23.4%	
Objectification	416	27.2%		43	25.1%	
Sexual-violence	197	12.9%		16	9.4%	
Misogyny-non-sexual-violence	164	10.7%		16	9.4%	

Table 3.2: EXIST2024 Label Distribution (Binary and Fine-grained).

3.2 Data Preprocessing and Feature Extraction

Multimodal Feature Integration To capture the multimodal nature of memes, both textual and visual information are incorporated:

Text Processing Both the MAMI and EXIST2024 datasets provide pre-extracted meme text. For EXIST2024, the organizers mention that images were "analyzed with an OCR software to extract the text"(Plaza et al., 2024), while the specific text extraction methodology for MAMI is not detailed in the available documentation. In this study, the provided meme text is used directly without additional preprocessing steps, maintaining the original formatting and linguistic characteristics as prepared by the dataset creators.

Image Caption Generation Visual content is processed using the BLIP-2 vision-language model, specifically the version incorporating FlanT5_XL fine-tuned on COCO (Li et al., 2023), to generate descriptive captions. These captions serve as textual representations of the visual content, enabling the analysis of linguistic features across both modalities.

Multimodal Integration For the ablation analysis employed in this work, meme text and image captions are concatenated. Different models adopt different integration strategies:

- **SVM**: Meme text and image caption are concatenated using a period separator ([meme text] . [image caption]).
- **BERT**: Meme text and image caption are concatenated following the sentence-pair encoding format by separating meme text and image caption with a special token [SEP], resulting in input of the form: [meme text] [SEP] [image caption].

For the example shown in Figure 3.3

Meme text: *"FIND THE CANADIAN FIND THE GIRL Without skin cancer. FUS VERY DEMOTIVATIONAL.com"*

Image caption: *"a group of women in bikinis standing next to a pool"*

SVM representation: *"FIND THE CANADIAN FIND THE GIRL Without skin cancer. FUS VERY DEMOTIVATIONAL.com. a group of women in bikinis standing next to a pool"*

BERT representation: *"FIND THE CANADIAN FIND THE GIRL Without skin cancer. FUS VERY DEMOTIVATIONAL.com [SEP] a group of women in bikinis standing next to a pool"*

Linguistic Feature Categories

This study examines four categories of linguistic features:

1. **Sentiment Markers**
   Sentiment features are extracted using the NRC Lexicon (Mohammad and Turney, 2013) and grouped as *positive* and *negative* sentiment words.

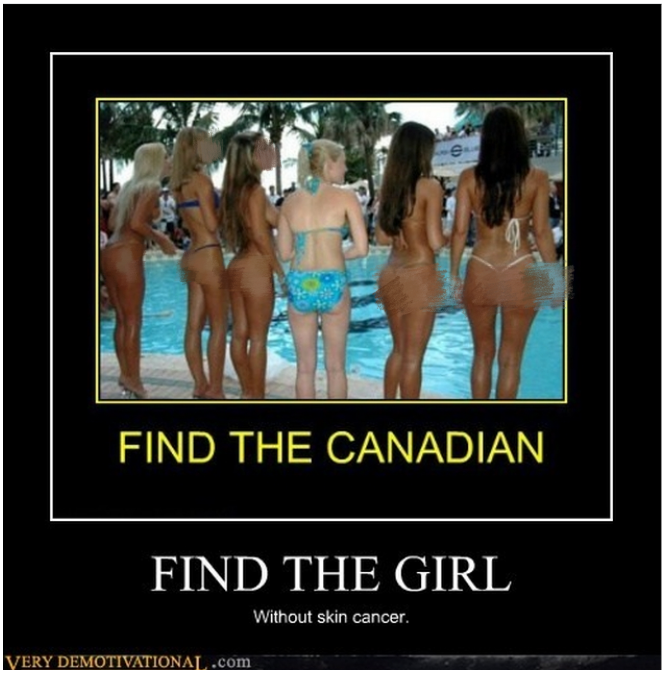


Figure 3.3: An Example of Meme from MAMI Dataset with ID: "10043.jpg".

2. **Emotion-based Features**
   Using the NRC Emotion Lexicon (Mohammad and Turney, 2013), words are categorized into eight emotion classes: *anger, anticipation, disgust, fear, joy, sadness, surprise, and trust*. Each emotion category is analyzed separately in the fine-grained multi-label classification tasks to determine its specific contribution to detecting different types of harmful content.

   Note: The vast majority of emotion words in the NRC Emotion Lexicon belong to either *positive* or *negative* sentiment categories. For example, the word *abandon* is labeled as *fear*, *sadness*, and also as *negative*. Therefore, there is substantial overlap between emotion-based and sentiment features.

3. **Function Words**
   Function words are categorized according to the Lancaster University stylistics framework (Leech et al., 2005) into nine closed-class grammatical categories: *determiners, pronouns, prepositions, conjunctions, auxiliary verbs, enumerators, particles, qualifiers, interjections* (shown in Table 3.3). The complete table is provided in Appendix 1.

Category	**Examples**
Determiners	*the, a, this, that, some, any, all*
Pronouns	*you, me, she, them, some, it, us*
Prepositions	*in, of, on, at, to, under, from*
Conjunctions	*and, but, or, if, then, although*
Auxiliary Verbs	*can, will, may, is, has, does, shall*
Enumerators	*one, three, first, second, eighteenth*
Particles	*no, not, up, out, off, down, about*
Qualifiers	*very, really, quite, rather, too*
Interjections	*oh, ah, ugh, hey, oops*

Table 3.3: Function Word Categories and Examples.

4. **Hate Speech Lexicon Terms**
   The HurtLex lexicon (Bassignana et al., 2018) is employed for its comprehensive categorization of offensive language across multiple languages. HurtLex contains 1,072 unique offensive, aggressive, and hateful words divided into 17 categories plus a macro-category indicating stereotype involvement. The distribution of entries varies significantly across categories, with the largest category being CDS (Derogatory words, 286 entries) and the smallest being RCI (Locations and demonyms, 9 entries). See Table 3.4 for the full list of categories and their respective entry counts.

Label	Description	Entries
PS	Negative stereotypes and ethnic slurs	41
RCI	Locations and demonyms	9
PA	Professions and occupations	81
DDF	Physical disabilities and diversity	22
DDP	Cognitive disabilities and diversity	64
DMC	Moral and behavioral defects	75
IS	Words related to social and economic disadvantage	16
OR	Plants	16
AN	Animals	107
ASM	Male genitalia	76
ASF	Female genitalia	30
PR	Words related to prostitution	54
OM	Words related to homosexuality	30
QAS	Words with potential negative connotations	79
CDS	Derogatory words	286
RE	Felonies and words related to crime and immoral behavior	35
SVP	Words related to the seven deadly sins of the Christian tradition	52
Total		**1,072**

Table 3.4: HurtLex Lexicon Categories, Descriptions, and Entry Counts.

3.3 Models

Support Vector Machine [1] (SVM)

The SVM model applies TF-IDF (Term Frequency-Inverse Document Frequency) weighted Bag-of-Words (BoW) vectorization to extract features from the combined textual input. This approach treats each word as an independent feature, creating a sparse vector representation where each dimension corresponds to a unique word in the vocabulary. TF-IDF weighting combines term frequency (how often a word appears in a document) with inverse document frequency (how rare the word is across the entire corpus), emphasizing words that are frequent within individual documents but rare across the collection. This method provides an effective way to capture the importance of linguistic features without considering word order or contextual relationships.

For input processing, meme text and image caption are concatenated using a period separator ([meme text] . [image caption]), creating a single unified text representation

[1] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

that captures both textual and visual information. Although the BoW representation does not preserve word order, this concatenation ensures that terms from each modality contribute jointly to the same feature space. The resulting combined text is then processed through TF-IDF vectorization to create the final feature representation for classification, where each word's importance is weighted according to its frequency.

BERT-based Model

The BERT model (Devlin et al., 2018) represents the deep learning approach, leveraging pre-trained contextual embeddings for more sophisticated text understanding. The model is built upon bert-base-uncased [2], a pre-trained transformer model with 12 layers, 768 hidden dimensions, and 12 attention heads. This architecture provides rich contextual representations that capture semantic relationships and dependencies between words, enabling the model to understand nuanced meanings that depend on context rather than just individual word presence.

For training, the model uses the following hyperparameters: 3 epochs, batch size of 16, learning rate of 2e-5 with linear learning rate scheduling, and a maximum sequence length of 128 tokens. The AdamW optimizer [3] is employed with linear warmup scheduling starting from 0 warmup steps. For multi-label classification, a hierarchical approach is implemented where binary classification is performed first, followed by fine-grained categorization applied only to instances classified as harmful in the binary stage.

Text is processed as a sentence pair using the format: [meme text] [SEP] [image caption]. This structure allows BERT to learn relationships between textual and visual semantic content through its attention mechanisms. The [SEP] token enables BERT to distinguish between the two modalities and potentially learn relationships across modalities. This approach leverages BERT's pre-trained understanding of sentence pair relationships, originally designed for tasks like natural language inference, to capture the interaction between meme text and visual content descriptions.

3.4 Ablation Study Methodology

Ablation Techniques

The systematic ablation study employs two primary techniques for each model architecture, designed to isolate the contribution of specific linguistic feature categories.

For SVM Models The ablation techniques for SVM focus on modifying the input text before BoW feature extraction:

- **Method A (Placeholder)**: Target words from specific feature categories are replaced with a generic placeholder token UNK. This method preserves the positional information and sentence structure while neutralizing the semantic content of the target features. The placeholder approach allows the model to recognize that something was present in that position but removes the specific linguistic information being tested.

  Example:

[2] https://huggingface.co/google-bert/bert-base-uncased
[3] https://docs.pytorch.org/docs/stable/generated/torch.optim.AdamW.html

  - Original: *EXTREME RAPE The woman in this picture is going to get it. Very hard. DIY.DESPAIR.COM. a woman with a tan shirt and a black belt*
  - Ablated: *EXTREME UNK The woman in this picture is going to get it. Very hard. DIY.DESPAIR.COM. a woman with a tan shirt and a black UNK*

- **Method B (Removal)**: Target words are completely removed from the input text, including any associated punctuation or spacing adjustments. This approach tests whether the mere presence of these words, regardless of their specific semantic content, contributes to classification performance.

  Example:

  - Original: *EXTREME RAPE The woman in this picture is going to get it. Very hard. DIY.DESPAIR.COM. a woman with a tan shirt and a black belt*
  - Ablated: *EXTREME The woman in this picture is going to get it. Very hard. DIY.DESPAIR.COM. a woman with a tan shirt and a black*

For BERT Models The ablation techniques for BERT leverage the model's built-in masking capabilities and contextual understanding:

- **Method A (Masking)**: Target words are replaced with [MASK] tokens, which are specifically designed for BERT's masked language modeling pre-training objective (Caselli et al., 2021). This approach is particularly suitable for BERT because the model has been trained to handle masked tokens and can potentially infer missing information from context.

  Example (ablating anger emotion words):

  - Original: *EXTREME RAPE The woman in this picture is going to get it. Very hard. DIY.DESPAIR.COM [SEP] a woman with a tan shirt and a black belt*
  - Ablated: *EXTREME [MASK] The woman in this picture is going to get it. Very hard. DIY.DESPAIR.COM [SEP] a woman with a tan shirt and a black [MASK]*

- **Method B (Removal)**: Target words are completely omitted from the input sequence, creating a more natural text flow without placeholder tokens. This method tests BERT's ability to maintain classification performance when specific linguistic features are entirely absent.

  Example (ablating anger emotion words):

  - Original: *EXTREME RAPE The woman in this picture is going to get it. Very hard. DIY.DESPAIR.COM [SEP] a woman with a tan shirt and a black belt*
  - Ablated: *EXTREME The woman in this picture is going to get it. Very hard. DIY.DESPAIR.COM [SEP] a woman with a tan shirt and a black*

Comparative Analysis

Using two approaches for each model allows for comprehensive understanding of feature importance. Placeholder/masking methods test whether semantic neutralization affects performance, while removal methods test whether the structural presence of features matters. Comparing results between these methods provides insights into whether models rely on specific semantic content or broader linguistic patterns.

However, it is important to note that direct comparison between SVM and BERT ablation results may not be entirely accurate due to fundamental differences in how these models process information. While SVM extracts features directly from input data, BERT relies on pre-trained representations - meaning that removing essential linguistic elements can disrupt the transfer learning process itself. The removal of content-bearing words may affect BERT's capacity to leverage its pre-trained knowledge (Caselli et al., 2021).

Experimental Procedure

The ablation study follows a structured approach designed to systematically identify the most influential linguistic features for binary and multi-label harmful content detection.

Binary Classification Analysis Initial experiments focus on binary classification (misogynous vs. non-misogynous for MAMI; sexist vs. non-sexist for EXIST2024) to identify feature categories that significantly impact performance. Four coarse feature types are evaluated: positive sentiment words, negative sentiment words, function words, and hate speech lexicon terms.

If ablation results in notable performance drops, indicating that the ablated features are important for classification, the study proceeds to fine-grained experiments for those specific feature categories.

Conversely, if marginal performance drops are observed, suggesting that the model does not heavily rely on the ablated features, Part-of-Speech (POS) ablation (Van Nooten et al., 2021) is conducted to understand what linguistic elements the model actually depends on when making predictions. The POS ablation focuses specifically on the six open-class word classes (shown in Table 3.5) from the Universal POS tag set (Universal Dependencies Consortium, 2024), as these represent the primary content-bearing elements of language. Each open-class POS category is systematically ablated to determine which grammatical categories the models rely on most heavily when linguistic features under consideration (positive/negative sentiment words, function words, and hate speech lexicon terms) show minimal impact on performance.

POS Tag	Category	Description / Examples
NOUN	Nouns	Nouns are a part of speech typically denoting a person, place, thing, animal or idea, e.g., *girl, car, tree*
VERB	Verbs	Verbs signal events and actions and often associated with grammatical categories like tense, mood, aspect and voice, e.g., *run, walked, thinking*
ADJ	Adjectives	Adjectives modify nouns and specify their properties or attributes, e.g., *beautiful, extreme, black*
ADV	Adverbs	Adverbs modify verbs for such categories as time, place, direction or manner, e.g., *quickly, very, here, extremely*
PROPN	Proper Nouns	Proper Nouns is a noun that is the name of a specific individual, place, or object, e.g., *Mary, Google, America*
INTJ	Interjection	Interjection is used most often as an exclamation or part of an exclamation., e.g., *hello, ouch*

Table 3.5: Universal Dependencies Open-Class POS categories.

Note: In the binary ablation stage, the feature categories are grouped into four coarse classes: positive sentiment words, negative sentiment words, function words, and hate speech lexicon terms. Emotion-based features are not separately tested at this stage because the vast majority of emotion words in the NRC lexicon fall under either positive or negative polarity. If the broader polarity groups (positive/negative sentiment words) are not indicative of performance differences, there would be no need to analyze the contribution of specific types of emotion words within these polarities.

Only if a feature group demonstrates notable performance drops in binary classification will it be further decomposed and examined in the fine-grained classification analysis. For instance, the hate speech lexicon, which is initially tested as a single group, is later disaggregated into 17 fine-grained subtypes (e.g., RE, CDS, etc.) to assess their category-specific effects.

Multi-label Classification Analysis First, I perform ablation experiments on four coarse feature categories (positive sentiment words, negative sentiment words, function words, and hate speech lexicon terms) to determine whether these broader linguistic

classes influence the model's ability to detect specific types of harmful content.

For those categories that show notable impact at the multi-label level, I then conduct fine-grained ablation to identify which specific features within each category (e.g., individual emotions like anger or sadness within the negative sentiment class, or specific function word types such as pronouns or auxiliaries) are most indicative of particular harm subtypes (e.g., objectification, stereotyping, or violence). This hierarchical approach enables a more detailed examination of how different linguistic features contribute to the detection of nuanced misogynistic and sexist phenomena.

Baseline

For each experiment, a comparison approach is employed to ensure reliable measurement of feature importance. Baseline performance is established using complete text without any ablation to provide reference metrics for comparison. Target features are then systematically replaced/ masked or removed according to the predetermined ablation techniques for each model. Performance comparison between baseline and ablated versions quantifies the impact of specific feature categories, with larger performance drops indicating higher feature importance for the classification task.

3.5 Evaluation Framework

3.5.1 Evaluation Metrics

Binary Classification Metrics

For binary classification tasks, macro-averaged precision, recall, and F1-score are employed to provide balanced assessment across both classes. While the MAMI test set is balanced (50% misogynous, 50% non-misogynous) and accuracy would be sufficient, I use consistent evaluation metrics across both datasets since EXIST2024 is imbalanced (56.4% sexist, 43.6% non-sexist). Additionally, F1-score for the positive class (sexist/misogynous content) is specifically reported to evaluate the model's effectiveness in detecting harmful content, which is the primary objective of this study. The macro-averaging approach ensures that both classes contribute equally to the overall performance assessment, preventing bias toward the majority class.

Multi-label Classification Metrics

For multi-label classification, this study adopts a hierarchical evaluation approach: multi-label classification is only considered for instances that are first classified as harmful in the binary stage. In this setup, multi-label performance is evaluated conditionally, ensuring that models are only penalized for misclassifications among relevant examples. This mimics real-world moderation scenarios, where resources are allocated to detailed categorization only after an initial binary classification step determines whether the content is harmful. Multi-label classification is implemented using the Binary Relevance [4] strategy, which transforms the multi-label problem into multiple independent binary classification tasks. For each fine-grained category (stereotyping, objectification, shaming, and violence in MAMI; or ideological-inequality, stereotyping-dominance, objectification, sexual-violence, and misogyny-non-sexual-violence in EXIST2024). This

[4] http://scikit.ml/api/skmultilearn.problem_transform.br.html

approach allows each category to be learned independently, though it does not capture potential dependencies between labels.

Macro-averaged precision, recall, and F1-score are reported to assess overall performance across all harm categories. In addition, per-category metrics (precision, recall, F1-score) are reported for specific subtypes of harmful content. This granularity enables identification of which harm categories are most affected by the ablation of specific linguistic features, while the hierarchical structure ensures that these scores are interpreted in the context of prior binary classification decisions. Specifically, fine-grained classification errors may stem from either binary classification failures (where harmful content is initially misclassified as non-harmful and never reaches the fine-grained stage) or from actual fine-grained categorization errors. This distinction is crucial for understanding whether feature ablation affects the initial detection of harmful content or the subsequent categorization of already-identified harmful instances.

3.5.2 Comprehensive Performance Assessment

Together, these metrics provide a comprehensive evaluation framework for both global and category-specific performance. This assessment design is crucial for understanding not just overall effectiveness, but also where models struggle within specific linguistic or multimodal contexts.

Chapter 4

Results

This chapter presents the results of systematic ablation experiments designed to evaluate the impact of specific linguistic feature categories on the detection of misogynous and sexist memes. The experiments cover both binary and fine-grained multi-label classification tasks, using two models (SVM and BERT) across two datasets (MAMI and EXIST2024).

The analysis begins with binary classification experiments involving the ablation of four broad feature categories (positive sentiment words, negative sentiment words, function words, and hate speech lexicons). Given that these coarse-level ablations produce minimal performance drops, the next section focuses on POS ablation experiments to probe more granular linguistic effects. This is followed by multi-label classification experiments using the same set of coarse feature categories. Based on the observed performance drops from coarse feature category ablation, fine-grained multi-label ablation is then applied to specific linguistic subtypes within the impactful coarse feature classes, revealing how the removal of individual features affects detection across different subcategories of misogynistic and sexist content. The results are complemented by a detailed error analysis to uncover failure patterns and identify the specific linguistic cues most critical for detecting each type of harmful content.

4.1 Binary classification

This section presents an examination of binary classification performance for harmful content detection across misogyny (MAMI dataset) and sexism (EXIST2024 dataset) detection tasks. The investigation employs systematic feature ablation methodologies to understand which linguistic components contribute most substantially to model performance, comparing traditional machine learning approaches (SVM) with modern transformer-based models (BERT). Through careful ablation studies examining sentiment words, hate speech lexicons, function words, and part-of-speech categories, this analysis reveals notable insights into the linguistic strategies employed by different models and the distinct patterns characterizing misogynous versus sexist content.

4.1.1 Feature Coverage and Distribution Analysis

Before examining ablation effects, I analyzed the distribution and coverage of linguistic features across both datasets to understand the scope of ablation experiments.

Feature Type	Avg Words/Doc (Train)	Avg Words/Doc (Test)	Number of Words (Train/Test)
Positive Sentiment	0.86	0.67	10,348 / 807
Negative Sentiment	0.70	0.51	8,779 / 616
Hate Speech Lexicons	1.26	1.02	12,982 / 1,043
Function Words	10.74	8.95	115,018 / 9,583

Table 4.1: MAMI Dataset Feature Distribution.

Feature Type	Avg Words/Doc (Train)	Avg Words/Doc (Test)	Number of Words (Train/Test)
Positive Sentiment	0.80	0.82	1,525 / 164
Negative Sentiment	0.68	0.70	1,286 / 138
Hate Speech Lexicons	1.13	1.28	1,791 / 224
Function Words	11.29	11.33	18,733 / 2,105

Table 4.2: EXIST2024 Dataset Feature Distribution.

Tables 4.1 and 4.2 present the feature distribution for MAMI and EXIST2024 datasets. They reveal consistent patterns across both datasets. MAMI demonstrates higher absolute feature coverage due to its larger dataset size, while EXIST2024 shows slightly higher per-document feature density in several categories. The coverage percentages indicate comprehensive feature representation and other categories showing substantial but more selective coverage ranging from 36.8% to 66.47% depending on feature type and dataset.

There are slight differences in feature statistics between SVM and BERT baseline experiments. It can be attributed to text preprocessing variations between models. BERT representations utilize cleaned text where BERT-specific tokens are removed ("[SEP]" tokens replaced with spaces), resulting in the format "Milk Milk.zip a pitcher of milk and a piece of cheese", whereas SVM representations maintain original punctuation formatting as "Milk Milk.zip. a pitcher of milk and a piece of cheese". This tokenization difference affects word boundary detection and feature extraction, leading to minor variations in average words per document and standard deviation measures while preserving overall feature distribution patterns and the reliability of experimental results.

Distribution of (Ablated) Words: Meme Text vs. Image Captions

To understand the differential contribution of textual and visual components in meme understanding, I analyzed the distribution of linguistic features between meme text and image captions generated by BLIP-2 vision-language model across both datasets.

Feature Type	Number of Words in Text	Number of Words in Caption	Text/Caption Ratio
Positive Sentiment	8,867	2,288	3.88:1
Negative Sentiment	8,043	1,352	5.95:1
Hate Speech Lexicons	8,498	4,709	1.80:1
Function Words	64,194	32,420	1.98:1

Table 4.3: MAMI Dataset Feature Distribution Between Meme Text and Image Captions.

Feature Type	Number of Words in Text	Number of Words in Caption	Text/Caption Ratio
Positive Sentiment	1,394	295	4.73:1
Negative Sentiment	1,245	179	6.96:1
Hate Speech Lexicons	1,156	674	1.72:1
Function Words	11,112	4,908	2.26:1

Table 4.4: EXIST2024 Dataset Feature Distribution Between Meme Text and Image Captions.

Tables 4.3 and 4.4 reveal distinct patterns in how linguistic features manifest across textual and visual modalities in meme content. Several key observations emerge from this comparative analysis:

Sentiment feature asymmetry can be observed from Tables 4.3 and 4.4. Both positive and negative sentiment features exhibit a strong textual predominance, with text-to-caption ratios ranging from 3.88:1 to 6.96:1 across datasets. Notably, negative sentiment features display the highest disparity (5.95:1 in MAMI, 6.96:1 in EXIST2024), suggesting that emotionally negative expressions are particularly concentrated in meme text rather than in image captions. One possible explanation for this pattern is the limited length and descriptive nature of BLIP-2-generated captions, which often omit subtle or affective linguistic cues that may be present in the original visual content.

Hate speech lexicon terms, by contrast, show the most balanced distribution between modalities, with relatively lower text-to-caption ratios (1.80:1 in MAMI, 1.72:1 in EXIST2024). This may be due to the implicit nature of harmful memes, which tend to avoid explicitly offensive vocabulary in the text. At the same time, BLIP-2-generated captions are descriptive and rarely contain offensive language, which further contributes to the observed balance.

Dataset Consistency: The remarkable consistency in distribution patterns between MAMI and EXIST2024 datasets validates the robustness of these findings across different linguistic contexts. The parallel trends in text-to-caption ratios across feature types suggest that the observed patterns are likely to represent fundamental characteristics of meme communication rather than dataset-specific artifacts.

Caption Generation Limitations and Their Impact on Feature Distribution

The substantially higher concentration of word features in text compared to image captions may be attributed to the inherent limitations of BLIP-2 model's caption generation approach. The BLIP-2 model tends to produce simple, descriptive captions that focus on basic visual elements and object identification rather than capturing nuanced emotional or contextual information.

Representative examples of BLIP-2-generated captions include:

- "a woman hugging a man in a comic strip" (MAMI)
- "a person holding a nokia cell phone" (MAMI)
- "a man and woman with a remote control" (EXIST2024)
- "two people sitting next to each other at a basketball game" (EXIST2024)

These captions demonstrate BLIP-2 models's preference for factual descriptions over emotionally nuanced language. This characteristic explains the limited amount of sentiment and especially hate speech features in the visual modality while maintaining relatively higher coverage of functional linguistic elements. The descriptive nature of BLIP-2-generated captions, while useful for basic scene understanding, might fail to capture the subtle emotional hints, ironic expressions, or socially suggestive elements often present in meme images. This results in the observed asymmetric feature distribution between textual and visual modalities.

4.1.2 Baseline Performance

As shown in Table 4.5, the SVM baseline models achieved macro-averaged F1-scores of 66% for **MAMI** and 64% for **EXIST2024**, establishing the performance baseline for subsequent ablation studies. These baseline results demonstrate moderate but consistent performance across both datasets, with the SVM model achieving slightly higher performance on MAMI than on EXIST2024. This difference may be attributed to the substantially larger training dataset size (10,000 vs 1,530 memes).

As presented in Table 4.6, the BERT baseline models achieved macro-averaged F1-scores of 71.4% for **MAMI** and 66.3% for **EXIST2024**, demonstrating improvements over SVM baselines (+5.4% and +2.3% respectively). These results establish BERT's superior contextual understanding capabilities across both datasets, with MAMI showing more pronounced gains than EXIST2024, consistent with empirical research demonstrating that transformer language model performance improves as dataset size increases, with larger datasets enabling more effective scaling and generalization (Kaplan et al., 2020).

On the MAMI dataset, BERT demonstrates enhanced discrimination capabilities, achieving balanced performance with high precision (74.9%) and recall (72.2%) for both classes (shown in Table 4.6). Notably, the misogynous class demonstrates strong recall (88.6%) with moderate precision (66.7%). This indicates BERT's sensitivity to harmful content detection, while non-misogynous content shows high precision (83%) but lower recall (55.8%), reflecting a conservative classification approach.

Results obtained on the EXIST 2024 dataset (shown in Table 4.6) show more modest improvements over SVM, with BERT achieving solid performance across both classes. The sexist class demonstrates balanced precision-recall characteristics (70.2%

and 69.5% respectively), while non-sexist content maintains consistent performance (62.3% precision, 63.2% recall). The smaller performance gap compared to MAMI suggests that BERT's advantages may be constrained by limited training data availability in sexism detection tasks.

4.1.3 Impact of Ablation Methods

SVM model

As can be observed in Table 4.5, the systematic comparison across both datasets and all feature categories demonstrated negligible variations between *placeholder/masking* method and *complete word removal* method.

Dataset	Feature Type	Placeholder Macro-F1 (%)	Remove Macro-F1 (%)	Difference (%)
MAMI	Baseline	66.0%		–
	Function Words	65.1%	64.5%	0.6%
	Hate Speech Lexicons	64.7%	64.9%	-0.2%
	Negative Sentiment	64.2%	64.2%	0.0%
	Positive Sentiment	66.8%	67.0%	-0.2%
EXIST2024	Baseline	64.0%		–
	Function Words	62.4%	61.3%	1.1%
	Hate Speech Lexicons	63.7%	62.7%	1.0%
	Negative Sentiment	62.1%	62.1%	0.0%
	Positive Sentiment	62.7%	62.5%	0.2%

Table 4.5: Impact of Placeholder vs. Remove Methods on SVM Binary Classification Performance (Macro-F1 in Percentages).

The consistently small variations across all conditions, with a maximum difference of 1.1% (function words for EXIST2024), indicate that the specific ablation technique has minimal impact on overall model performance. This finding justifies the decision to focus on the removal method for subsequent analyses due to its computational efficiency.

BERT Model

For the BERT model, the comparison (shown in Table 4.6) between *masking methods* and *complete word removal* revealed more substantial variations compared to the SVM model, with differences ranging from -5.9% to 3.6%. These larger variations in BERT performance might be attributed to the model's contextual understanding capabilities. When words are masked with [MASK] tokens, BERT can leverage its pre-trained masked language modeling objective to infer contextual relationships, potentially maintaining some semantic information. Conversely, complete removal creates gaps in the input sequence that may disrupt BERT's attention mechanisms differently. This can lead to varied impacts on classification performance depending on the linguistic features being ablated.

4.1.4 Feature Ablation vs. Random Word Control

To establish whether observed performance drops genuinely reflect feature importance rather than general vocabulary reduction, I compared targeted feature ablation against

Dataset	Feature Type	Mask Macro-F1 (%)	Remove Macro-F1 (%)	Difference (%)
MAMI	Baseline	71.4%		–
	Function Words	69.1%	66.6%	2.5%
	Hate Speech Lexicons	70.2%	68.6%	1.6%
	Negative Sentiment	67.8%	68.0%	-0.2%
	Positive Sentiment	69.0%	70.9%	-1.9%
EXIST2024	Baseline	66.3%		–
	Function Words	62.7%	68.6%	-5.9%
	Hate Speech Lexicons	73.2%	69.6%	3.6%
	Negative Sentiment	70.1%	69.5%	0.6%
	Positive Sentiment	69.9%	68.6%	1.3%

Table 4.6: Impact of Mask vs. Remove Methods on BERT Binary Classification Performance (Macro-F1 in Percentages).

random word ablation for content-bearing categories (positive sentiment, negative sentiment, and hate speech lexicons) . The random ablation strategy targeted content words only, excluding function words from the random selection pool. This control method matched the exact number of words ablated in each document during feature ablation, randomly removing the same quantity of content words from identical sentences. This approach ensures a fair comparison by controlling for both vocabulary reduction volume and document-level impact.

Dataset	Feature Type	Feature Ablation (%)	Random Ablation (%)	Feature vs. Baseline (%)	Random vs. Baseline (%)
MAMI	Baseline	66.0%		–	
	Hate Speech Lexicons	64.9%	65.9%	-1.1%	-0.1%
	Negative Sentiment	64.2%	65.4%	-1.8%	-0.6%
	Positive Sentiment	67.0%	65.5%	+1.0%	-0.5%
EXIST2024	Baseline	64.0%		–	
	Hate Speech Lexicons	62.7%	63.7%	-1.3%	-0.3%
	Negative Sentiment	62.1%	61.8%	-1.9%	-2.2%
	Positive Sentiment	62.5%	67.2%	-1.5%	+3.2%

Table 4.7: Feature vs. Random Ablation: Macro F1 Score Comparison Against Baseline (in Percentages).

The comparison (Table 4.7) reveals striking patterns when examining performance drops relative to baseline, where larger drops indicate greater feature importance. Hate speech lexicon features demonstrate modest discriminative value, with targeted ablation causing larger performance drops than random controls in both datasets (MAMI: 1.1% vs 0.1% drop; EXIST2024: 1.3% vs 0.3% drop). Negative sentiment features show dataset-dependent importance, with MAMI exhibiting substantial discriminative value (1.8% feature drop vs 0.6% random drop) while EXIST2024 shows minimal difference between targeted and random ablation (1.9% vs 2.2% drops).

The most counterintuitive patterns emerge with positive sentiment features, which contradict intuitive expectations that removing positive sentiment indicators might improve harmful content detection. In MAMI, removing positive sentiment words actually improves performance above baseline (+1.0%), while random ablation causes a perfor-

mance decrease (-0.5%). This suggests positive words may introduce classification noise. EXIST2024 presents a more complex pattern where feature ablation causes a modest drop (-1.5%) but random ablation paradoxically improves performance (+3.2%). This indicates that positive sentiment words in sexism detection contexts may play multiple, potentially conflicting roles. They may be used ironically in sexist content or be present in ambiguous cases where positive language masks underlying sexist attitudes, but they might also help models distinguish non-sexist examples by providing cues typically associated with positive language.

These findings reveal that the linguistic features examined in this study provide limited discriminative advantage over random vocabulary ablation. To better understand which linguistic categories drive model performance, the analysis proceeded to examine part-of-speech ablation.

4.1.5 Part-of-Speech Ablation Analysis

Given the marginal effects observed in feature-specific ablation above, I conducted systematic POS ablation to identify which word categories most significantly impact binary classification performance. Due to time constraints, this analysis focused on open-class words as defined by the Universal Dependencies framework (de Marneffe et al., 2021), which include content words that carry the primary semantic meaning in sentences.

Methodology

Using the removal method, I systematically ablated six open-class POS categories (NOUN, VERB, ADJ, ADV, PROPN, and NUM) to assess performance drops caused by ablation across two different models: traditional SVM classifiers and transformer-based BERT models. This approach reveals more interpretable patterns than traditional feature-based approaches and allows for comparative analysis of linguistic feature importance across different models.

Ablation Statistics

To provide insight into the linguistic composition of the datasets and the extent of feature removal during ablation, I presented detailed statistics for each POS category across both models and datasets.

The statistics (shown in Table 4.8 and Table 4.9) reveal several important patterns in the linguistic composition of our datasets. Nouns consistently represent the largest lexical category across both datasets, with approximately 6-6.5 words per document on average. Proper nouns show variation between the two datasets, with MAMI containing significantly more proper noun references (2.7-3.5 words per document) compared to EXIST2024. This difference may reflect the distinct nature of the content, with misogynous content potentially containing more references to specific individuals, organizations, or entities. The distribution of verbs also shows variation for two datasets, with EXIST2024 containing slightly higher verb density than MAMI. Adjectives and adverbs maintain relatively consistent distributions across datasets, while interjections remain the least frequent category, appearing in fewer than 15% of documents across all conditions.

POS Category	Avg Words/Doc (Train)	Avg Words/Doc (Test)	Total Words Removed (Train/Test)
Noun	6.29	5.04	62,855 / 5,038
Proper Noun	3.35	2.67	33,472 / 2,669
Verb	2.82	2.21	28,234 / 2,206
Adjective	1.28	0.96	12,848 / 963
Adverb	0.65	0.53	6,488 / 526
Interjection	0.13	0.10	1,283 / 98

Table 4.8: MAMI Dataset POS Category Distribution.

POS Category	Avg Words/Doc (Train)	Avg Words/Doc (Test)	Total Words Removed (Train/Test)
Noun	6.40	6.56	9,787 / 1,122
Verb	3.00	3.09	4,588 / 529
Proper Noun	2.94	3.20	4,505 / 547
Adjective	1.48	1.41	2,272 / 241
Adverb	0.77	0.68	1,172 / 117
Interjection	0.17	0.25	253 / 42

Table 4.9: EXIST2024 Dataset POS Category Distribution.

The minor variations in POS category statistics between SVM and BERT models can be attributed to text preprocessing variations between models, as discussed in Section [4.1.1] for feature statistics. These preprocessing differences affect word boundary detection and consequently impact POS tagging accuracy, leading to minor differences in the category distributions while preserving overall distribution patterns and experimental reliability.

POS Ablation Results

Tables 4.10 and 4.11 present the POS ablation results for both SVM and BERT models, revealing both consistent patterns and model-specific sensitivities across misogyny and sexism detection tasks.

Consistent Patterns Across SVM and BERT Models Several linguistic patterns emerge consistently across both models. Nouns demonstrate substantial importance, especially for MAMI dataset, causing 4.7-5.0% F1 drops in SVM models and 9.9% in BERT for misogyny detection. This consistent impact reflects nouns' role as primary carriers of semantic content, particularly in misogynous contexts where categorical references and derogatory terminology are prevalent. Additionally, this pattern may be influenced by nouns being the most frequent lexical category in both datasets (Tables 4.8 and 4.9).

Proper nouns appear to show relatively strong discriminative power across both models, with what seem to be particularly pronounced effects in sexism detection (6.5%

Dataset	POS Category	Resulting Macro-F1 (%)	Macro-F1 Drop (%)
MAMI	Baseline	66.0%	–
	Noun	61.2%	4.7%
	Verb	64.6%	1.4%
	Proper Noun	64.7%	1.2%
	Adjective	65.7%	0.3%
	Interjection	66.5%	-0.5%
	Adverb	66.6%	-0.6%
EXIST2024	Baseline	64.0%	–
	Proper Noun	57.5%	6.5%
	Noun	59.1%	5.0%
	Adjective	61.1%	2.9%
	Adverb	63.9%	0.1%
	Interjection	64.2%	-0.2%
	Verb	65.2%	-1.2%

Table 4.10: SVM Model POS Ablation: Macro-F1 Score Impact per Category.

F1 drop in SVM, 3.8% in BERT for EXIST2024 datasets). This pattern suggests that named entities and specific references may constitute important features for identifying sexist content, potentially indicating their relevance regardless of the underlying classification approach.

A dataset-specific pattern emerges with verb ablation, demonstrating consistent behavior across both models. In misogyny detection contexts, verb removal consistently degrades performance across both SVM (1.4% drop) and BERT (4.1% drop) models, indicating that action-oriented language plays a meaningful role in identifying misogynous content. However, in sexism detection tasks, despite verbs being the second most frequent category in EXIST2024 (Table 4.9), verb removal consistently improves performance in both SVM (-1.2% drop) and BERT (-3.9% drop) models. This convergent finding across different models suggests that verbs may systematically introduce classification noise in sexism detection, possibly because everyday action-oriented language patterns may mask the subtle linguistic markers that distinguish sexist from non-sexist content.

Model-Specific Differences The most striking model-specific pattern emerges in BERT's differential treatment of nouns across datasets. While SVM models show relatively consistent noun importance across both tasks (4.7% drop in misogyny, 5.0% drop in sexism), BERT exhibits a sensitivity contrast—devastating impact in misogyny detection (9.9% drop) versus virtually no impact in sexism detection (-0.2% drop, indicating slight improvement). This dramatic 10.1 percentage point difference represents the largest difference between the two tasks observed in the POS ablation analysis.

This stark contrast suggests that BERT may have learned fundamentally different linguistic strategies for each classification task. The model appears to rely heavily on nominal structures and categorical references for misogyny detection, where specific groups are often targeted through derogatory terminology. Conversely, for sexism detection, BERT seems to have developed classification strategies that are largely in-

Dataset	POS Category	Resulting Macro-F1 (%)	Macro-F1 Drop (%)
MAMI	Baseline	71.4%	–
	Noun	61.5%	9.9%
	Proper Noun	64.1%	7.3%
	Verb	67.3%	4.1%
	Adjective	67.9%	3.5%
	Interjection	70.1%	1.3%
	Adverb	71.3%	0.1%
EXIST2024	Baseline	66.3%	–
	Proper Noun	62.5%	3.8%
	Adjective	66.1%	0.2%
	Adverb	66.3%	0.0%
	Interjection	66.3%	0.0%
	Noun	66.5%	-0.2%
	Verb	70.2%	-3.9%

Table 4.11: BERT Model POS Ablation: Macro-F1 Score Impact per Category.

dependent of nominal content, possibly focusing on syntactic patterns, implicit biases, or contextual cues that transcend specific word categories.

Adjectives also reveal interesting differences between models in their discriminative contribution. While SVM models show consistent moderate importance in sexism detection (2.9% F1 drop) with minimal impact in misogyny tasks (0.3% drop), BERT maintains stronger adjective sensitivity in misogyny detection (3.5% drop) with reduced importance in sexism contexts (0.2% drop). This reversal suggests that the two models may capture different aspects of evaluative and descriptive language patterns.

Another interesting pattern emerges when comparing SVM and BERT sensitivity across classification tasks. For the MAMI dataset, BERT exhibits greater sensitivity to POS ablations than SVM (e.g., noun ablation: 9.9% vs. 4.7%). However, this trend is reversed in the EXIST2024 dataset, where SVM often shows greater sensitivity than BERT (e.g., proper noun ablation: 6.5% in SVM vs. 3.8% in BERT). One plausible explanation for this inversion lies in dataset size: MAMI contains substantially more training samples than EXIST2024, enabling the context-aware BERT model to generalize more robustly. In contrast, in lower-resource settings such as EXIST2024, BERT may overfit to specific lexical signals and become less dependent on certain POS categories, leading to diminished ablation effects.

Low-Frequency Category Effects The minimal impact of adverb and interjection removal across both models (typically less than 1.5% F1 change) can be attributed to their low frequency in the datasets, with fewer than 1 word per document on average. This frequency-performance relationship demonstrates consistent behavior across both SVM and BERT models, indicating that lexical category importance correlates with occurrence frequency regardless of model complexity. To verify whether dataset size indeed accounts for this inversion, future work could replicate the experiment using a smaller subset of MAMI that matches the size of EXIST2024.

Comparative Analysis and Discussion The comparative analysis between SVM and BERT models provides insights into how different modeling approaches leverage linguistic features for misogynous and sexist content detection. The observed patterns suggest several implications for both model selection and understanding of harmful language detection tasks.

BERT's heightened sensitivity to lexical category removal (with noun ablation performance drops reaching 9.9% compared to SVM's maximum 6.5%) may reflect its contextual processing capabilities, which appear to create stronger dependencies on specific word categories. The sensitivity to nouns is particularly notable given that they are the most frequent lexical category in both datasets (Tables 4.8 and 4.9). This increased sensitivity also presents a trade-off: while BERT achieves superior baseline performance, it also makes BERT more sensitive to input changes, especially when specific types of words are intentionally modified or omitted.

The divergent task-specific patterns across models, where SVM prioritizes proper nouns for sexism detection while BERT emphasizes nouns for misogyny detection, suggest that different models may capture complementary linguistic signals. Rather than indicating different strategies employed by harmful content creators, these patterns may reflect the varying capacity of different models to detect subtle linguistic markers within similar content types.

The consistent performance drops when removing certain categories (particularly verbs in sexism detection) across both models provide more robust evidence for potential classification noise. However, the magnitude differences between models (SVM: -1.2%, BERT: -3.9%) indicate that contextual BERT models may be more susceptible to such interference, possibly due to their ability to form complex associations that incorporate various linguistic patterns.

4.2 Multi-Label Classification Performance

Building upon the binary classification analysis, this section examines the more challenging task of multi-label classification. In this task, models must simultaneously identify specific subtypes of misogynous/sexist content. This differs from binary classification, which simply distinguishes between misogynous/sexist and non-misogynous/sexist text.

The multi-label approach provides granular insights into different manifestations of misogyny and sexism. This fine-grained classification task reveals the complex linguistic landscape of gender-based harmful content. It also highlights the severe challenges posed by class imbalance and data sparsity in minority categories.

The analysis begins with a systematic comparison of SVM and BERT baseline performance across multiple content categories. This comparison establishes the relative effectiveness of traditional SVM models versus transformer-based BERT models in handling multi-class classification complexity. Subsequently, I conducted detailed linguistic feature ablation studies. These studies focus on sentiment indicators, emotional expression, function words, and hate speech lexicons at both broad and fine-grained levels. The goal is to uncover the distinct linguistic strategies that characterize different forms of harmful content. Due to time constraints, the fine-grained multi-label ablation analysis is conducted exclusively on the SVM model.

4.2.1 Baseline Performance

SVM Model

The SVM multi-label classifiers achieved macro-averaged F1-scores of 47.6% on MAMI (Table 4.12) and 40.4% on EXIST2024 (Table 4.13), demonstrating the increased complexity of multi-label classification compared to binary tasks and revealing substantial dataset-specific challenges. Both datasets exhibit significant class imbalance issues, though the imbalance appears differently in each dataset.

The MAMI results reveal distinct patterns across misogyny categories (Table 4.12). Among misogynous categories, objectification demonstrated the most balanced performance (F1=58.6%) with comparable precision and recall values around 60%. In contrast, shaming presented the greatest classification challenge with the lowest F1-score (32.7%). Violence showed an interesting pattern with high precision (64.5%) but substantially lower recall (32.0%), indicating the model's difficulty in identifying all violent content instances.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Non-misogynous	79.6%	46.8%	58.9%	500
Shaming	35.8%	30.1%	32.7%	146
Stereotype	42.6%	47.1%	44.8%	350
Objectification	60.8%	56.6%	58.6%	348
Violence	64.5%	32.0%	42.8%	153
Micro avg	57.2%	46.0%	51.0%	1497
Macro avg	56.7%	42.5%	47.6%	1497
Weighted avg	60.8%	46.0%	51.3%	1497
Samples avg	55.9%	47.7%	49.7%	1497

Table 4.12: Results for SVM for fine-grained classes on MAMI.

EXIST2024 presented more severe classification challenges and results varied more between classes as we can see from Table 4.13. Among sexist categories, ideological-inequality and objectification showed relatively stable performance (F1=53.7% and 50.0% respectively). However, violence-related categories demonstrated significant weaknesses. Misogyny-non-sexual-violence achieved the lowest performance (F1=17.4%) with extremely low recall (12.5%), while sexual-violence exhibited precision-recall imbalance (50.0% precision vs. 25.0% recall), similar to the violence pattern observed in MAMI.

Both datasets consistently show that violence-related categories present classification difficulties, particularly in recall performance. This pattern suggests that violent content may require more sophisticated feature representations or additional training data to achieve reliable detection. The precision-recall imbalances observed across both datasets indicate that bag-of-words representations may not adequately capture the nuanced semantic patterns that distinguish different forms of misogynous and sexist content.

Class	Precision	Recall	F1-Score	Support
Non-sexist	61.8%	55.3%	58.3%	76
Ideological-inequality	51.2%	56.4%	53.7%	39
Stereotyping-dominance	27.7%	32.5%	29.9%	40
Objectification	46.9%	53.5%	50.0%	43
Sexual-violence	50.0%	25.0%	33.3%	16
Misogyny-non-sexual-violence	28.6%	12.5%	17.4%	16
Micro avg	47.7%	46.1%	46.9%	230
Macro avg	44.3%	39.2%	40.4%	230
Weighted avg	48.1%	46.1%	46.4%	230
Samples avg	49.2%	48.3%	47.5%	230

Table 4.13: Results for SVM for fine-grained classes on EXIST2024.

BERT Model

The BERT multi-label classifiers show mixed performance compared to SVM baselines. BERT achieves macro-averaged F1-scores of 52.4% on MAMI (Table 4.14) and 35.1% on EXIST2024 (Table 4.15). BERT demonstrates clear improvement on MAMI with a notable 4.8% macro F1 gain over SVM. However, it exhibits a concerning 5.3% F1 decline on EXIST2024.

Class	Precision	Recall	F1-Score	Support
Non-misogynous	86.9%	53.0%	65.8%	500
Shaming	46.1%	32.2%	37.9%	146
Stereotype	50.5%	45.7%	48.0%	350
Objectification	62.3%	53.2%	57.4%	348
Violence	72.7%	41.8%	53.1%	153
Micro avg	65.0%	48.2%	55.3%	1497
Macro avg	63.7%	45.2%	52.4%	1497
Weighted avg	67.2%	48.2%	55.7%	1497
Samples avg	66.2%	51.9%	55.9%	1497

Table 4.14: Results for BERT for fine-grained classes on MAMI.

BERT shows consistent improvements across most MAMI categories (Table 4.14). Among misogynous categories, objectification demonstrates the most robust performance (F1=57.4%) with balanced precision-recall characteristics, similar to the SVM pattern. Violence classification exhibits high precision (72.7%) but substantially lower recall (41.8%), resulting in F1=53.1%. This precision-recall imbalance mirrors the SVM results but with improved overall performance.

Shaming remains the most challenging category (F1=37.9%) despite improvements over SVM, consistent with the difficulty in detecting subtle forms of psychological manipulation. Stereotype detection achieves moderate performance (F1=48.0%), suggesting that stereotypical language patterns remain challenging even for transformer-based models.

EXIST2024 presents more severe challenges for BERT, with the 35.1% macro-

Class	Precision	Recall	F1-Score	Support
Non-sexist	74.6%	61.8%	67.6%	76
Ideological-inequality	52.2%	61.5%	56.5%	39
Stereotyping-dominance	36.5%	47.5%	41.3%	40
Objectification	44.4%	46.5%	45.5%	43
Sexual-violence	0.0%	0.0%	0.0%	16
Misogyny-non-sexual-violence	0.0%	0.0%	0.0%	16
Micro avg	53.4%	47.8%	50.5%	230
Macro avg	34.6%	36.2%	35.1%	230
Weighted avg	48.2%	47.8%	47.6%	230
Samples avg	51.8%	51.9%	50.4%	230

Table 4.15: Results for BERT for fine-grained classes on EXIST2024.

averaged F1 representing a decline from SVM performance (Table 4.15). Among sexist categories, ideological-inequality demonstrates the strongest performance (F1=56.5%) with moderate precision (52.2%) and recall (61.5%). Stereotyping-dominance shows moderate performance (F1=41.3%) but with lower precision (36.5%) than recall (47.5%).

Most critically, both sexual-violence and misogyny-non-sexual-violence categories achieve complete failure, with precision, recall, and F1-score all equal to zero. This represents a more severe degradation than observed with SVM models and indicates complete inability to detect these critical forms of sexist content.

The contrasting performance between datasets suggests that BERT's effectiveness is heavily dependent on training data characteristics. The MAMI improvements may be attributed to the larger training dataset (10,000 vs 1,530 memes) and BERT's enhanced ability to capture contextual semantic patterns in meme-based content. However, the EXIST2024 deterioration, particularly the complete failure on sexual-violence (197 training instances, 12.9% of the training data) and misogyny-non-sexual-violence (164 training instances, 10.7% of the training data) categories, indicates that BERT may require substantially larger datasets to learn robust representations for rare but critical categories in fine-grained classification tasks.

4.2.2 Linguistic Feature Ablation

This section examines the impact of different linguistic feature sets on multi-label classification performance. The analysis is conducted at both a broad and a fine-grained level. Due to time constraints, the fine-grained multi-label ablation analysis is conducted exclusively on the SVM model.

- **Coarse feature classes (for SVM and BERT)**: Four broad linguistic feature categories (positive sentiment, negative sentiment, function words, and hate lexicon) are individually removed to assess their overall impact across subcategories.

- **Fine-grained feature groups (for SVM only)**: For coarse categories showing notable impacts, individual linguistic features (e.g., emotion_sadness, hate_cds) are further analyzed to identify which specific features affect the detection of different subtypes of misogynous and sexist content.

The coarse feature classes ablation reveals which broad feature categories are important, while the fine-grained analysis explains which specific linguistic elements within these broad categories are especially impactful for subcategory-level predictions.

Coarse Feature Classes

SVM Model: Tables 4.16 and 4.17 present the macro and per-category F1-scores after coarse-grained feature ablations using the SVM classifier in multi-label setting.

Across both the MAMI and EXIST2024 datasets, the removal of negative sentiment features causes the largest macro-averaged F1 degradation, with macro-F1 drops of 3.5 percentage points on MAMI and 4.2 on EXIST2024. This confirms the critical role of emotionally negative language in distinguishing harmful categories from non-harmful content.

In the MAMI dataset, negative sentiment removal leads to substantial F1 drops in the shaming (–6.3%) and violence (–6.0%). Similarly, in EXIST2024, negative sentiment removal has a marked impact on ideological-inequality (–5.4%), sexual-violence (–7.2%), and misogyny-non-sexual-violence (–8.7%) categories.

Interestingly, hate speech lexicon features, though often associated with more explicit toxic language, produce the second-largest performance drops across both datasets. This result is somewhat counterintuitive given the subtle nature of harmful memes.

Lastly, in both datasets, their removal of function words consistently reults in performance drop in detection of objectification (MAMI: –3.5%; EXIST2024: –1.8%). This is likely because such words help establish factors that are especially important in identifying objectifying language such as who is acting, who is being described, and from what point of view.

Method	Non-misogyny	Shaming	Stereotype	Objectification	Violence	Macro Avg
Baseline	58.9%	32.7%	44.8%	58.6%	42.8%	47.6%
Pos_Sentiment (Remove)	60.7%	34.0%	45.2%	58.7%	39.3%	47.6%
Neg_Sentiment (Remove)	57.6%	26.4%	43.3%	56.5%	36.8%	44.1%
Hate Lexicons (Remove)	57.8%	32.1%	43.9%	57.6%	37.9%	45.8%
Function Words (Remove)	57.3%	30.3%	45.7%	55.1%	43.3%	46.4%

Table 4.16: MAMI Dataset Results on SVM Model: F1 Scores for Different Categories and Macro-averaged.

Method	Non-sexist	Ideological inequality	Stereotyping dominance	Objectification	Sexual violence	Misogyny-non sexual-violence	Macro Avg
Baseline	58.3%	53.7%	29.9%	50.0%	33.3%	17.4%	40.4%
Pos_Sentiment (Remove)	55.7%	48.2%	34.0%	49.5%	30.8%	16.7%	39.1%
Neg_Sentiment (Remove)	55.9%	48.3%	29.8%	48.4%	26.1%	8.7%	36.2%
Hate Lexicons (Remove)	56.3%	53.0%	31.2%	52.1%	8.7%	17.4%	36.5%
Function Words (Remove)	56.4%	51.8%	33.3%	48.2%	37.0%	17.4%	40.7%

Table 4.17: EXIST2024 Dataset Results on SVM Model: F1 Scores for Different Categories and Macro-averaged.

BERT Model: Tables 4.18 and 4.19 report the multi-label F1 scores of the BERT classifier after the removal of coarse feature classes.

For the MAMI dataset, the removal of negative sentiment features causes the largest performance degradation in macro-F1 (–3.4%), particularly affecting the shaming (–3.1%) and violence (–11.5%) categories. Hate speech lexicon removal also yields noticeable performance drops (macro-F1: –2.1%), with larger reductions in the violence category.

In the EXIST2024 dataset, the most substantial macro-F1 drop is observed after removing function words (–3.7%), with complete prediction failures (F1 = 0.0%) in the sexual-violence and misogyny-non-sexual-violence categories. Interestingly, negative sentiment removal results in performance gains in categories like ideological-inequality, but causes a drop on sexual-violence.

Method	Non-misogynous	Shaming	Stereotype	Objectification	Violence	Macro Avg
Baseline	64.4%	33.7%	44.1%	56.4%	52.2%	50.2%
Pos_Sentiment (Remove)	64.1%	33.2%	43.5%	57.6%	55.1%	50.7%
Neg_Sentiment (Remove)	60.0%	30.6%	43.8%	58.8%	40.7%	46.8%
Hate Lexicons (Remove)	59.6%	32.7%	46.1%	54.8%	47.5%	48.1%
Function Words (Remove)	60.8%	31.9%	43.7%	56.2%	53.2%	49.2%

Table 4.18: MAMI Dataset Results on BERT Model: F1 Scores for Different Categories and Macro-averaged.

Method	Non-sexist	Ideological inequality	Stereotyping dominance	Objectification	Sexual violence	Misogyny-non sexual-violence	Macro Avg
Baseline	62.0%	55.0%	37.0%	56.0%	11.1%	0.0%	36.9%
Pos_Sentiment (Remove)	63.4%	53.7%	48.1%	46.9%	11.1%	0.0%	37.2%
Neg_Sentiment (Remove)	68.0%	63.3%	46.3%	54.7%	0.0%	0.0%	38.7%
Hate Lexicons (Remove)	62.5%	60.2%	45.2%	43.0%	20.0%	0.0%	38.5%
Function Words (Remove)	59.9%	56.5%	40.7%	42.0%	0.0%	0.0%	33.2%

Table 4.19: EXIST2024 Dataset Results on BERT Model: F1 Scores for Different Categories and Macro-averaged.

The results suggest the hypothesized balancing effects of individual feature ablations. Removing certain features can benefit some subcategories while harming others, resulting in relatively small macro-F1 changes.

Fine-grained Feature Analysis

Based on the results from the coarse feature classes ablation, where negative sentiment, hate speech lexicons, and function word features showed impact, this section investigates which specific fine-grained features within those broader categories contribute to subcategory-level classification performance in both the MAMI and EXIST2024 datasets.

Table 4.20 lists the top 10 feature types that lead to the largest macro-F1 score drops when ablated from the model input.

Emotion Features **MAMI Dataset**: Table 4.21, we can see that the removal of individual emotion features results in notable changes across specific misogyny subcategories. Features associated with the emotion sadness have the most substantial impact on Shaming (F1 drop from 32.7% to 24.9%, a decrease of 7.8%) and Violence (42.8% to 34.9%, 7.9% drop), which may indicate that emotionally charged language contributes meaningfully to the classification of these subtypes. Disgust features similarly affect Shaming (32.7% to 26.3%, 6.4% drop) and Violence (42.8% to 37.9%, 4.9% drop), possibly reflecting the role of degrading language in these categories. Fear shows a selective influence, especially on Violence (drop of 6.8%), while Anger affects Violence to a lesser extent (drop of 5.9%). In contrast, Objectification and Stereotype demonstrate more stable F1 scores under feature ablation, suggesting that these subcategories might rely less on explicit emotional content and more on other linguistic cues.

Dataset	Category	Macro-F1 after removal (%)	Macro-F1 Drop (%)
MAMI	Baseline	47.6%	–
	emotion_sadness	43.7%	3.9%
	emotion_disgust	44.9%	2.6%
	emotion_fear	45.2%	2.4%
	emotion_anger	45.7%	1.9%
	function_pronouns	46.6%	0.9%
	hate_pr	47.1%	0.5%
	hate_re	47.3%	0.2%
	function_particles	47.4%	0.2%
	hate_dmc	47.4%	0.1%
	function_auxiliary	47.4%	0.1%
EXIST2024	Baseline	40.4%	–
	emotion_anger	37.3%	3.1%
	function_interjections	38.2%	2.3%
	hate_cds	38.7%	1.7%
	function_auxiliary	39.0%	1.4%
	emotion_disgust	39.2%	1.3%
	function_particles	39.3%	1.1%
	function_prepositions	39.3%	1.1%
	hate_om	39.3%	1.1%
	hate_re	39.4%	1.1%
	emotion_sadness	39.9%	0.6%

Table 4.20: SVM Model Feature Ablation: Macro-F1 Score Impact per Category.

EXIST2024 Dataset: Table 4.22 present subcategory F1 scores for negative emotion feature removal for EXIST2024 Dataset. Among these emotion features, anger has the most pronounced effect across multiple sexism subcategories. Its removal leads to the largest observed performance drop in Misogyny-non-sexual-violence (F1 decreases from 17.4% to 8.0%, a 9.4% drop), followed by a substantial reduction in Sexual-violence (33.3% to 27.3%, 6.0% drop) and a moderate decline in Objectification (50.0% to 47.3%, 2.7% drop). Notably, Stereotyping-dominance shows a counterintuitive increase (29.9% to 32.6%) after anger features are removed. This provides support for balancing effects mentioned above - ablation improves performance in some subcategory while degrading others. Disgust features have moderate influence on Ideological-inequality (drop of 3.1%) and Objectification (drop of 2.2%), with smaller but measurable impact on Sexual-violence (1.3% decrease). Fear demonstrates a similar pattern, primarily influencing Sexual-violence (drop of 1.3%) and marginally Misogyny-non-sexual-violence (17.4% to 16.7%, 0.7% drop). Sadness produces its clearest effect in Stereotyping-dominance (29.9% to 28.3%, 1.6% drop), while having little impact on the other subcategories. These results suggest that emotional cues contribute selectively across sexism categories.

Hate Speech Lexicon Features **MAMI Dataset**: From Table 4.23, it is noticed that the removal of prostitution-related terms (hate_pr) leads to a slight decrease in F1

Feature	Non-misogyny	Shaming	Stereotype	Objectification	Violence
Baseline	58.9%	32.7%	44.8%	58.6%	42.8%
Anger (Remove)	58.3%	31.1%	44.2%	57.9%	36.9%
Disgust (Remove)	58.1%	26.3%	44.0%	58.4%	37.9%
Fear (Remove)	57.3%	30.6%	43.8%	58.0%	36.0%
Sadness (Remove)	56.9%	24.9%	43.7%	57.9%	34.9%

Table 4.21: MAMI Dataset - Subcategory F1 Scores for Fine-grained Negative Emotion Feature Removal.

Feature	Non-sexist	Ideological-inequality	Stereotyping-dominance	Objectification	Sexual-violence	Misogyny-non-sexual-violence
Baseline	58.3%	53.7%	29.9%	50.0%	33.3%	17.4%
Anger (Remove)	57.3%	51.2%	32.6%	47.3%	27.3%	8.0%
Disgust (Remove)	56.3%	50.6%	30.8%	47.8%	32.0%	17.4%
Fear (Remove)	58.3%	52.5%	29.2%	52.7%	32.0%	16.7%
Sadness (Remove)	55.9%	53.7%	28.3%	50.5%	33.3%	17.4%

Table 4.22: EXIST2024 Dataset - Subcategory F1 Scores for Fine-grained Negative Emotion Feature Removal.

for Stereotype (from 44.8% to 43.5%, a 1.3% drop) and Violence (from 42.8% to 42.1%, a 0.7% drop), suggesting that these lexical cues carry marginal discriminative value for certain misogynous expressions. Prior studies indicate that sexualized and commodifying language including prostitution-related slurs often serves as a key marker of misogynous discourse, especially in violent or threatening contexts (Njagi et al., 2015). The observed impact on the Violence category aligns with these findings, as such terms may amplify aggression while reinforcing objectification. Crime-related terms (hate_re) show a more pronounced effect, especially on Violence, where F1 drops from 42.8% to 39.1%, representing a 3.7% decline. Interestingly, their removal results in an improvement in Shaming (F1 increases from 32.7% to 34.0%, a 1.3% gain), indicating that these features may introduce ambiguity or noise for some subcategories while remaining essential for others. The effect of moral and behavioral defect terms (hate_dmc) appears minor overall, with a drop in Shaming (32.7% to 31.7%, 1.0% decline), suggesting a limited role in classification for this subcategory.

Feature	Non-misogyny	Shaming	Stereotype	Objectification	Violence
Baseline	58.9%	32.7%	44.8%	58.6%	42.8%
DMC (Remove)	59.2%	32.5%	44.4%	57.7%	43.3%
PR (Remove)	58.4%	33.8%	43.5%	57.8%	42.1%
RE (Remove)	59.5%	34.0%	46.0%	57.9%	39.1%

Table 4.23: MAMI Dataset - Subcategory F1 Scores for Fine-grained Hate Speech Lexicon Feature Removal.

EXIST2024 Dataset: Table 4.24 shows derogatory terms (hate_cds) produce the strongest effects among hate lexicon features, particularly in the most severe sexism categories. Their removal results in a notable F1 drop in Sexual-violence (from 33.3% to 25.0%, an 8.3% decline) and a moderate decline in Misogyny-non-sexual-violence (17.4% to 14.8%, 2.6% drop). Other minority-related terms (hate_om) selectively impact Stereotyping-dominance, where performance drops from 29.9% to 27.7% (a 2.2% decline). Crime-related terms (hate_re) also play a notable role, especially in detecting

Sexual-violence, where removal leads to a 7.2% performance loss (F1 drops from 33.3% to 26.1%). These results suggest that the sexist content in this dataset may rely more heavily on subtle or implicit expressions of hostility.

Feature	Non-sexist	Ideological-inequality	Stereotyping-dominance	Objectification	Sexual-violence	Misogyny-non-sexual-violence
Baseline	58.3%	53.7%	29.9%	50.0%	33.3%	17.4%
CDS (Remove)	55.3%	53.7%	32.6%	50.5%	25.0%	15.4%
OM (Remove)	56.3%	53.0%	27.7%	51.1%	32.0%	16.0%
RE (Remove)	57.3%	54.3%	32.3%	49.5%	26.1%	16.7%

Table 4.24: EXIST2024 Dataset - Subcategory F1 Scores for Fine-grained Hate Speech Lexicon Feature Removal.

Function Word Features The investigation of function words is motivated by research demonstrating that stylometric features indicate "the existence of a linguistic register for hate speech messages with specific stylistic properties and a negative emotional load," with focusing on these features leading to "more cross-domain robustness" (Markov et al., 2021).

MAMI Dataset: Among function word categories, pronouns exert the most noticeable impact (shown in Table 4.25). Their removal causes performance drops in Shaming (from 32.7% to 30.7%, a 2.0% decline) and Objectification (from 58.6% to 56.4%, a 2.2% decline), indicating that referential expressions play an important role in identity-targeted content. Prior work has shown that gendered language, including pronouns, contributes to gender bias by making gender more salient and reinforcing stereotypic views (Bigler and Leaper, 2015). This supports the idea that pronoun usage helps distinguish subtle forms of psychological manipulation and identity-based targeting characteristic of shaming behaviors. Interestingly, Violence shows improvement after ablating pronouns (F1 increases from 42.8% to 45.7%, a 2.9% gain), which may suggest that direct or impersonal language characterizes this category more strongly. Auxiliary verbs and particles demonstrate modest effects. Removing auxiliary verbs leads to a slight decrease in Violence (42.8% to 41.9%, 0.9% drop), while particles show mixed effects, including a small decline in Shaming (32.7% to 31.9%) and a gain in Violence (42.8% to 44.3%).

Feature	Non-misogyny	Shaming	Stereotype	Objectification	Violence
Baseline	58.9%	32.7%	44.8%	58.6%	42.8%
Auxiliary (Remove)	59.9%	33.0%	45.4%	57.1%	41.9%
Particles (Remove)	59.0%	31.9%	45.3%	56.5%	44.3%
Pronouns (Remove)	56.2%	30.7%	44.1%	56.4%	45.7%

Table 4.25: MAMI Dataset - Subcategory F1 Scores for Fine-grained Function Word Feature Removal.

EXIST2024 Dataset: Table 4.26 shows that interjections show the largest effect among function words. Their removal causes a notable drop in sexual-violence (F1 decreases from 33.3% to 25.0%, an 8.3% decline), with moderate impacts on other categories. Auxiliary verbs notably affect Objectification, where performance drops from 50.0% to 44.7% (a 5.3% decrease), while improving classification of Stereotyping-dominance (F1 increases from 29.9% to 34.0%, a 4.1% gain). Prepositions show mixed effects, producing a 2.4% gain in Stereotyping-dominance and a minor increase in Ob-

jectification. Particles also influence Sexual-violence (F1 drops from 33.3% to 30.8%, a 2.5% decline). Although pronouns do not appear among the top macro-F1 impact features, further examination reveals their removal reduces Objectification performance (50.0% to 46.7%) but significantly improves Sexual-violence classification (33.3% to 41.7%, an 8.4% increase). This suggests that pronoun use in sexist discourse may play diverging roles across subcategories. Prior research (Alorainy et al., 2019) has shown that pronouns are useful for distinguishing between in-group and out-group perspectives in hate speech detection, which could partly explain their role in capturing biased perspectives in sexist language.

Feature	Non-sexist	Ideological-inequality	Stereotyping-dominance	Objectification	Sexual-violence	Misogyny-non-sexual-violence
Baseline	58.3%	53.7%	29.9%	50.0%	33.3%	17.4%
Auxiliary (Remove)	59.7%	54.3%	34.0%	44.7%	24.0%	17.4%
Interjections (Remove)	53.9%	53.0%	30.4%	50.0%	25.0%	16.7%
Particles (Remove)	55.3%	53.7%	28.9%	50.5%	30.8%	16.7%
Prepositions (Remove)	54.9%	52.4%	32.3%	50.5%	33.3%	16.7%
Pronouns (Remove)	53.2%	54.1%	30.3%	46.7%	41.7%	18.2%

Table 4.26: EXIST2024 Dataset - Subcategory F1 Scores for Fine-grained Function Word Feature Removal.

Cross-Dataset Patterns and Balancing Effects Both datasets show that sadness-related features exert a measurable impact on the classification of non-misogynous or non-sexist content. In the MAMI dataset, removing sadness features reduces macro F1 in the Non-misogyny category from 58.9% to 56.9%, a 2.0% decline. In EXIST2024, the corresponding drop in Non-sexist classification is from 58.3% to 55.9%, a 2.4% decrease. These results suggest that the presence of sadness-related words helps the model recognize content as non-misogynous or non-sexist. When these cues are removed, the model is less able to distinguish non-harmful instances.

Ablating all negative emotion features leads to substantial performance drop in both datasets, which aligns with expectations. In MAMI, the largest drops appear in Shaming (from 32.7% to 24.9%, 7.8% decline) and Violence (from 42.8% to 34.9%, 7.9% decline). In EXIST2024, the steepest declines occur in Sexual-violence (from 33.3% to 26.1%, 7.2% drop) and Misogyny-non-sexual-violence (from 17.4% to 8.0%, 9.4% drop). These patterns confirm that emotional expressions are essential for detecting the most severe and overtly harmful categories across both datasets.

The results also provide empirical support for the hypothesized balancing effects of individual feature ablations. Removing certain features can benefit some subcategories while harming others, resulting in relatively small macro-F1 changes that may obscure substantial subcategory-specific shifts. For instance, in the MAMI dataset, removing pronouns leads to performance improvements in Violence (F1 increases from 42.8% to 45.7%, +2.9%), but causes declines in Shaming and Objectification (-2.0% and -2.2%, respectively). Similarly, in EXIST2024, ablating auxiliary verbs improves classification of Stereotyping-dominance (from 29.9% to 34.0%), but degrades performance in Objectification (from 50.0% to 44.7%) and Sexual-violence (from 33.3% to 24.0%). These trade-offs illustrate that certain linguistic features contribute differently to the representation of various subcategories, and their removal can create countervailing effects that complicate aggregate metric interpretation.

4.2.3 Error Analysis

To better understand the model's limitations and failure patterns, I conducted a two-part error analysis focusing on (1) POS ablation experiments for binary classification and (2) fine-grained features ablation experiments for multi-label classification. The coarse-grained binary ablation, targeting four categories negative sentiment words, positive sentiment words, function words, and hate speech lexicon terms, was intentionally excluded from error analysis, as its overall performance degradation was minimal and provided limited insight into model behavior.

For the POS ablation analysis, 20 false negatives (FN) and 20 false positives (FP) were randomly selected from cases where the baseline model predicted correctly but the prediction became incorrect after ablation. For the fine-grained multi-label classification analysis, up to 10 false negative cases were examined for each category, depending on availability.

In the POS ablation analysis, I examined both false negatives (FN) and false positives (FP). This analysis was conducted on both the MAMI and EXIST2024 datasets using SVM and BERT classifiers. Based on prior POS ablation results, noun ablation (NOUN) in the MAMI dataset led to the highest performance drops, while proper noun ablation (PROPN) was particularly impactful in the EXIST2024 dataset. These configurations are thus selected for in-depth error analysis.

In contrast, the multi-label ablation analysis focused solely on false negatives (FN) in order to investigate why the removal of specific word types caused the model to misclassify misogynous/sexist memes as non-misogynous/non-sexist. This analysis was performed using the SVM model under a remove-ablation setting, where specific semantic features (e.g., emotion_sadness, hate_cds, func_pronoun) were removed from the input representations. The goal was to identify which features are essential for correctly detecting specific fine-grained categories and to understand why their removal leads to model failure.

By analyzing examples of both FN and FP errors across these two ablation regimes, I demonstrate how syntactic and semantic features play distinct yet complementary roles in identifying nuanced forms of harmful content. These insights clarify which linguistic cues are most critical for maintaining classification accuracy and inform the design of future systems.

The following sections provide detailed breakdowns for each dataset (MAMI and EXIST2024), highlighting the categories most affected by specific feature removals and offering illustrative examples to support key findings.

POS Ablation

MAMI Dataset Ablation of NOUN (the most frequent lexical category in the MAMI dataset) revealed substantial error increases. Nouns in this dataset often encode gendered slurs (e.g., *bitch*, *tits*), identity references (e.g., *woman*, *girl*), or physical descriptors tied to visual dehumanization. Their removal eliminates key referents that allow models to anchor harmful discourse.

False Negative Example:
Meme Text: "IF IT'S GOT TITS OR TIRES IT WILL GIVE YOU PROBLEMS"
Caption: "a woman in a wheelchair"
Fine-grained: ['shaming', 'stereotype', 'objectification']
Ablated Words: ['tits', 'woman', 'wheelchair']

In this case, removing key nouns like *tits* and *woman* stripped the model of both physical and gender markers that define objectification. Although the remaining text preserves the syntactic structure, its ideological content becomes opaque, leading to misclassification.

Conversely, noun ablation also induced false positives. These arose when benign memes lost neutral or humorous objects (e.g., *pig*, *hamburger*), causing residual phrases to appear more emotionally intense or sarcastic. This suggests that when such neutral nouns are removed, the classifier may misinterpret the harmless content as harmful.

False Positive Example:
Meme Text: "WHAT IF I'M REALLY ATTRACTIVE AND HOT GIRLS JUST THINK I'M OUT OF THEIR LEAGUE"
Caption: "a man with a tie"
Ablated Words: ['i'm', 'girls', 'league', 'man', 'tie']

In this example, ablation removed key nouns and pronouns such as *girls*, *league*, and *man*. These terms grounded the meme in a humorous and self-deprecating social commentary. Once removed, the remaining fragments *"WHAT IF REALLY ATTRACTIVE AND HOT JUST THINK OUT"* appeared emotionally ambiguous or even self-aggrandizing, prompting the model to misclassify the meme as misogynous.

EXIST2024 Dataset Despite proper nouns (PROPN) ranking third in frequency among lexical categories in EXIST2024, the ablation of proper nouns triggered frequent misclassifications, especially in categories ideological-inequality and sexual-violence. Proper nouns like *Feminism*, *Daddy*, or references to social institutions encode ideological frames essential for interpreting sexist intent. Their removal neutralizes anchoring cues, leading to reduced classifier sensitivity to rhetorical sexism.

False Negative Example:
Meme Text: "This is not Feminism. This is Feminism."
Caption: "a woman holding a sign that says this is not feminist"
Fine-grained: ['ideological-inequality']
Ablated Words: ['Feminism', 'feminist']

Here, removing ideological terms such as *Feminism* entirely defuses the critique embedded in the meme. Without these terms, the model fails to recognize the contrastive framing strategy that underlies this meme's sexist messaging.

A separate but related source of error stems from the widespread use of full-uppercase formatting in EXIST memes. Such formatting amplifies perceived emotional intensity. When content-bearing nouns or named entities are removed, what remains is often short, capitalized exclamatory text. This textual imbalance exaggerates the classifier's emotional signal, resulting in false positives.

False Positive Example:
Meme Text: "WHEN IT COMES TO SEXUAL ORIENTATION, MANY PEOPLE FORGET THAT THE PEOPLE THEY ARE BULLYING OR JUDGING"
Caption: "a girl with a group of people in front of her"
Ablated Words: ['sexual', 'orientation', 'human', 'too', 'gender', 'discrimination', 'workplace']

This case reveals a critical issue with POS tagging accuracy. The full-uppercase formatting caused the POS tagger to misclassify many common words as proper nouns. As a result, the proper noun ablation removed substantially more content than intended, including crucial terms such as *sexual, orientation*, and *discrimination*. The resulting fragmented input lacked key contextual cues, leading the model to misclassify the meme as harmful, therefore resulting in a false positive.

Multi-label Ablation

MAMI Dataset Shaming & Violence errors were primarily triggered by the removal of emotionally negative words(sadness, digust, fear and anger), such as *fat*, *rape*, and *bitch*. These terms, though lexically simple, serve as strong cues for derogatory or violent content. Their absence weakened the classifier's ability to recognize implicit hostility or objectifying rhetoric, especially in memes with otherwise mild or ambiguous context.

False Negative Example (Shaming):
Meme Text: "I'M NOT SAYING I'M FAT I JUST HAVE MORE CHINS THAN A ASIAN PHONE BOOK"
Caption: "a woman with a fat face and a phone"
Ablated Words (disgust): ['fat', 'fat']
Fine-grained Label: ['shaming', 'stereotype']
Baseline Predictions: ['misogynous', 'shaming', 'stereotype']
Ablation Predictions: ['non-misogynous']
This meme was originally correctly classified as both shaming and stereotype. This means that the model was able to detect the offensive message targeting body appearance. After removing the emotional word linked to negative judgment *'fat'*, the model misclassified the meme as non-misogynous. This suggests that the model relies on obvious and emotional terms to recognize harmful content. Once these were removed, the rest of the sentence was not strong enough for the model to recognize it as offensive.

False Negative Example (Violence):
Meme Text: "WOMEN NEED A GOOD BEATING ONCE IN A WHILE memegenerator.net"
Caption: "a man with a beard and a beard"
Ablated Words (fear): ['beating']
Fine-grained Label: ['objectification', 'violence']
Baseline Predictions: ['misogynous', 'violence']
Ablation Predictions: ['misogynous', 'objectification']
Removing the word *beating* stripped the meme of its overt physical aggression cue. Although objectifying elements remained, the absence of this explicit violence-related term caused the classifier to fail to detect the the violence label.

Additionally, functional pronouns (e.g., *she*, *you*, *my*) played a critical role in shaming and objectification categories. Their removal caused failures in identifying subject-object relationships and personal targeting, which are essential for detecting gendered aggression or sexualized framing. For instance, without pronouns, a meme may lose clear referential structure, reducing the model's confidence in assigning a misogynistic label.

False Negative Example (Shaming):
Meme Text: "SHE WAS SAYING SOME SHIT ABOUT HOW HER DAY WENT Sandtman AND ALL I HEARD WAS MY NAUGHTY ASS THOUGHTS"

Caption: "a woman in a dress"
Ablated Words (pronoun): ['she', 'some', 'her', 'all', 'i', 'my']
Fine-grained Label: ['stereotype', 'objectification']
Baseline Predictions: ['misogynous', 'stereotype', 'objectification']
Ablation Predictions: ['non-misogynous']
After the pronouns were removed, the sentence no longer clearly indicated who was talking, whose day was being talked about, and whose thoughts were being expressed. Without these personal references, the model could not recognize that the speaker was judging the woman and turning her into a sexual object. As a result, it failed to detect the objectifying tone embedded in the male gaze and misclassified the meme as non-misogynous.

False Negative Example (Objectification):
Meme Text: "I DONT ALWAYS STEAL YOUR WIFE RIE BUT WHEN I DO, I MAKE HER MY SEX SLAVE makoorems.ca"
Caption: "a man sitting at a table with a beer"
Ablated Words (pronoun): ['i', 'your', 'i', 'i', 'her', 'my']
Fine-grained Label: ['objectification', 'violence']
Baseline Predictions: ['misogynous', 'stereotype', 'objectification', 'violence']
Ablation Predictions: ['misogynous', 'stereotype', 'violence']

While the meme retains violent connotations through phrases like *make her my sex slave*, it simultaneously conveys sexual objectification by reducing the woman to a passive, possessed entity. The term *sex slave* is semantically dual-purpose: it implies physical coercion (violence) and sexual commodification (objectification). Therefore, ablation of pronouns like *her* and *my* disrupted the possessive framing, leading the model to miss the objectification cue.

This pattern highlights a key limitation: while emotion-related terms provide direct signals of hostility, pronouns function as syntactic and pragmatic anchors that shape who is speaking, who is the target of the hate speech, and how aggression is framed. The model's reliance on both indicates its sensitivity to both lexical and discourse-level cues in multi-label meme classification.

EXIST2024 Dataset Based on the multilabel ablation analysis previously conducted on the EXIST2024 dataset, anger-related words ablation caused widespread performance drops across nearly all sexist subcategories except *stereotyping-dominance*. The following examples illustrate how the removal of such cues led to false negatives (FN) across ideological, objectifying, and violent expressions.

False Negative Example (Ideological-Inequality):
Meme Text: "Proud Walmart. Astocate The economy is going great! We created 5 million jobs this year Cash y tax check"
Caption: "a woman talking to a man in a store..."
Ablated Words (anger): ['cash']
Fine-grained Label: ['ideological-inequality', 'stereotyping-dominance']
Baseline Predictions: ['sexist', 'ideological-inequality', 'stereotyping-dominance']
Ablation Predictions: ['sexist', 'stereotyping-dominance', 'objectification']

Although *cash* appears lexically neutral, it anchors a sarcastic critique of economic disparity and institutional inequality. The phrase contrasts economic success with in-

ability to pay rent, implying systemic economic injustice—core to ideological-inequality. Its removal disrupted this ideological signal, resulting in the classifier overlooking the injustice critique and misclassifying the meme.

False Negative Example (Objectification):
Meme Text: "Sure, women like money and Jewelry. But what makes them fall in love is when you... GRAB BOTH BUNS & TEAT IT LIKE A MAN..."
Caption: "a billboard with a large advertisement for a hamburger..."
Ablated Words (anger): ['money', 'grab']
Fine-grained Label: ['objectification', 'sexual-violence']
Baseline Predictions: ['sexist', 'objectification']
Ablation Predictions: ['non-sexist']

Here, the removal of *grab* and *money* erased cues tied to sexual commodification and economic stereotyping, undermining the model's ability to detect objectifying content. Without these cues, the sentence became structurally fragmented and lost the framing of women as sexual commodities, leading to failure in detecting objectification.

False Negative Example (Sexual-Violence):
Meme Text: "I CAN'T TALK RIGHT NOW grand theft auto I'M DOING HOT GIRL SH*T"
Caption: "a woman in a bikini taking a picture of herself"
Ablated Words (anger): ['theft', 'hot']
Fine-grained Label: ['objectification', 'sexual-violence']
Baseline Predictions: ['sexist', 'stereotyping-dominance', 'objectification', 'sexual-violence']
Ablation Predictions: ['sexist', 'stereotyping-dominance', 'objectification']

Terms like *theft* and *hot* contribute metaphorical framing—linking women to criminalized or hypersexual roles. The removal of such aggressive and suggestive terms diminished the model's sensitivity to the sexually exploitative undertones of the meme, resulting in a failure to flag sexual-violence.

False Negative Example (Misogyny-Non-Sexual-Violence):
Meme Text: "WOMEN AND MEN ARE CREATED EQUALLY, THAT MEANS EQUAL PAY, AND RIGHTS EXCEPT MEN CAN'T HIT WOMEN. THAT'S ILLEGAL"
Caption: "a woman with dreadlocks and a hat"
Ablated Words (anger): ['hit', 'illegal']
Fine-grained Label: ['ideological-inequality', 'misogyny-non-sexual-violence']
Baseline Predictions: ['sexist', 'ideological-inequality', 'stereotyping-dominance', 'misogyny-non-sexual-violence']
Ablation Predictions: ['sexist', 'ideological-inequality', 'stereotyping-dominance']

This meme presents a call for gender equality, but the final clause *"EXCEPT MEN CAN'T HIT WOMEN. THAT'S ILLEGAL"* introduces a hostile undercurrent. The phrase reframes legal protections against domestic violence as unfair treatment, implying resentment toward women's social advantages.

The terms *hit* and *illegal* are crucial in surfacing this hostility: they explicitly invoke physical aggression and encode a legal judgment. Without them, the meme reduces to a complaint about unequal standards, which aligns more closely with the *ideological-inequality* category. As a result, the classifier misses the underlying non-sexualized

misogynistic aggression.

These cases illustrate that anger-related terms do more than just express emotion. They also play a key role in shaping how harm is communicated, especially in indirect ways common in online sexist memes. Their removal impairs the model's ability to disambiguate ironic, sarcastic, or elliptical expressions of harm that are common in online sexist discourse.

False Negative Example (Objectification):
Meme Text: "if she stand like this she got no ass Sticks and stones might break my bones but thot poses will never hurt me headass bitches..."
Caption: "a girl in ripped jeans with a stick and stone in her butt..."
Ablated Words (pronouns): ['she', 'this', 'she', 'my', 'me', 'her']
Fine-grained Label: ['objectification']
Baseline Predictions: ['sexist', 'objectification']
Ablation Predictions: ['non-sexist']

Removing personal pronouns disrupted referential coherence and syntactic alignment between subject and evaluative predicate (e.g., *"she got no ass"*). This ambiguity weakened the model's ability to resolve female agency, ultimately preventing it from detecting objectification. Without pronouns like *she* and *her*, the insult appears more abstract, obscuring the targeted objectification that originally underpinned the meme. The classifier thus failed to capture its sexist framing.

Chapter 5

Discussion and Conclusions

5.1 Concluding Remarks

This thesis investigated which linguistic features are most indicative of harmful meme detection, with a particular focus on misogynous and sexist content across the MAMI and EXIST2024 datasets. The central research question addressed how different linguistic feature categories (*sentiment markers*, *emotion features*, *function words*, and *hate speech lexicon terms*) affect model performance across classification settings (binary vs. multi-label), models (SVM vs. BERT), and datasets.

Two models were used: a linear SVM with TF-IDF-based Bag-of-Words representations, and a BERT-based transformer model (bert-base-uncased). The feature ablation strategy followed a structured design. For **binary classification**, coarse-level (*sentiment markers*, *emotion features*, *function words*, and *hate speech lexicon terms*) ablation was first applied. If a feature category proved impactful, fine-grained ablation was conducted to examine the contribution of individual subcategories within that feature group; otherwise, part-of-speech (POS) ablation was used to explore which open-class word categories (*NOUN, VERB, ADJ, ADV, PROPN,* and *NUM*) drove classification performance. For **multi-label classification**, coarse-level ablation was also applied first, and fine-grained ablation followed only for categories that caused a substantial performance drop.

In binary classification, the removal of sentiment markers, function words, and hate speech lexicon terms led to minimal performance degradation across both datasets and models. This suggests that these features are not strong predictors for binary classification. To identify which word categories most significantly impact binary classification performance, POS ablation was conducted. In MAMI, removing **NOUN** (common nouns, the most frequent lexical category) led to the largest macro-F1 drop for both SVM and BERT models, highlighting the importance of content-bearing words in misogynistic meme detection. In EXIST2024, the largest drop was caused by removing **PROPN** (proper nouns), although they ranked only third in frequency. Error analysis suggested that this effect may be amplified by POS tagging errors, especially in all-uppercase meme text.

In the multi-label classification setting, ablation experiments revealed that different types of harmful content are tied to different linguistic features. **Sentiment markers** (including **emotion-related features**), particularly negative emotions, were among the most impactful. In MAMI, *negative sentiment* was the most impactful coarse feature class, especially for shaming (–6.3%) and violence (–6.0%). Fine-grained ab-

lation showed that all four negative emotion types contributed to these drops, with *sadness*-related words having the strongest effect (–7.8% for shaming, –7.9% for violence). In EXIST2024, *negative sentiment* again had the largest impact, particularly for ideological-inequality (–5.4%), sexual-violence (–7.2%), and misogyny-non-sexual-violence (–8.7%). Among negative emotion types, *anger*-related words were especially influential for detecting sexual-violence (–6.0%) and misogyny-non-sexual-violence (–9.4%) categories.

Hate speech lexicon terms had limited impact, reflecting the subtle and implicit nature of harmful meme expressions. The only notable effect came from the *CDS (derogatory terms)* subtype, which caused an 8.3% drop in F1 score for sexual-violence detection in EXIST2024.

Function words, although not content-bearing, showed several meaningful effects. *Pronouns* were important for detecting objectification in both datasets, likely because they help the model identify who is the target of the harmful content. Pronouns have also been proven to be useful for distinguishing between in-group and out-group references in hate speech detection. In MAMI, *pronouns* also contributed to detecting shaming. In EXIST2024, *interjections* had an unexpected but notable impact on sexual-violence detection, suggesting that informal emotional signals may be leveraged by the model in harmful meme contexts.

5.2 Limitations

While the findings presented in this thesis offer valuable insights into feature reliance in misogynous and sexist meme detection, several limitations constrain their generalizability and interpretability.

First, the scope of this study was restricted by time constraints, which limited the diversity of models, datasets, and feature categories that could be explored. In particular, fine-grained multi-label ablation was conducted only on the SVM pipeline, while BERT experiments were confined to binary classification and coarse-level multi-label classification and thus lack insight into category-specific dependencies.

Second, the BERT experiments did not include random-word ablation as a control, which limits our ability to attribute performance changes to specific linguistic feature categories. Since BERT processes contextualized token embeddings, the removal of high-frequency or content-bearing words could distort sentence meaning more than in feature-based models. This makes controlled comparison challenging.

Third, the use of POS ablation depends on accurate preprocessing and tagging. Given the prevalence of noisy formatting, non-standard grammar, and uppercase emphasis in memes, some POS assignments may be unreliable, thereby affecting ablation outcomes.

Fourth, as highlighted in Van Nooten et al. (2021), it is important to note that BERT and SVM operate on different learning assumptions. While SVM extracts features directly from the input and is thus highly interpretable under ablation, BERT's reliance on prior pretrained knowledge and contextual embeddings complicates interpretability. Consequently, feature removal in BERT not only removes lexical items but may also disrupt pretrained expectations, especially when meaningful content words are omitted. These model disparities should be considered when interpreting comparative results.

Finally, the quality of visual modality features was constrained by the descriptive

limitations of BLIP-2. BLIP-generated captions tended to emphasize literal scene content rather than capturing emotionally charged, culturally referential, or derogatory visual signals. This asymmetry likely reduced the model's ability to leverage multimodal inputs effectively.

5.3 Future Work

Future research could extend the findings of this thesis in several important directions. First, implementing fine-grained multi-label ablation studies using transformer-based architectures would shed light on how contextualized embeddings interact with specific linguistic categories. Introducing random-word control ablations into these models would further clarify whether observed performance drops stem from specific feature removal or general disruption.

Second, the specificity and semantic granularity of visual input remain critical limiting factors in multimodal meme detection. Future work should explore the integration of grounding-based captioning models, scene graph generation, or multimodal attention frameworks that more effectively align visual semantics with textual harms.

Third, new linguistic features such as rhetorical markers, syntactic templates, and figurative language (e.g., metaphors and irony) could provide richer cues for understanding subtle or covert expressions of harm. These higher-level features may be especially useful in detecting less explicit subtypes such as stereotyping-dominance or ideological-inequality.

Fourth, the ablation framework could be extended to multilingual contexts, especially in low-resource or culturally underrepresented languages. Doing so would help build more inclusive hate speech detection systems and evaluate the cross-linguistic generalizability of feature dependencies.

Lastly, future studies could explore two additional dimensions: (i) analyzing semantic relationships between text and image (e.g., reinforcement, contradiction, or neutralization) using Natural Language Inference (NLI) models, and (ii) examining the relationship between modality feature disparities and classification errors. These directions may offer further insight into multimodal inconsistency and help develop more robust cross-modal alignment strategies.

Appendix A - Complete Function Word Categories

Category	Complete Word List
Determiners	*the, a, an, this, that, these, those, my, your, his, her, its, our, their, any, each, every, some, all, both, either, neither, few, many, much, several, more, most, less, no, enough, which, what, whose*
Pronouns	*i, me, my, mine, myself, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, we, us, our, ours, ourselves, they, them, their, theirs, themselves, who, whom, whose, which, what, this, that, these, those, anybody, somebody, nobody, everybody, anyone, someone, no one, everyone, each, either, neither, one, all, some, many, few, any, none, both*
Prepositions	*of, at, in, on, without, between, under, over, beside, through, during, among, across, against, towards, around, before, after, along, behind, below, beyond, despite, except, from, inside, near, onto, outside, past, since, till, until, upon, within, about, above, beneath, beside, by, down, into, like, off, out, throughout, to, toward, underneath, unto, up, with, without, regarding, round*
Conjunctions	*and, but, or, nor, for, yet, so, although, because, since, unless, until, when, while, whereas, after, before, if, then, even though, though, as long as, provided that, however, therefore, thus, moreover, nevertheless*
Auxiliary Verbs	*am, is, are, was, were, be, being, been, have, has, had, having, do, does, did, will, would, shall, should, can, could, may, might, must, ought*
Enumerators	*one, two, three, four, five, six, seven, eight, nine, ten, first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth, once, twice, thrice*
Particles	*no, not, nor, as, to, up, out, off, down, about, around, aside, away, back, apart*
Qualifiers	*very, really, quite, somewhat, rather, too, pretty, fairly, slightly, almost, nearly, barely, hardly, scarcely, completely, absolutely, totally, utterly, extremely, especially, particularly, specifically*
Interjections	*oh, ah, ugh, hey, oops, gadzooks, wow, ouch, eh, hmm*

Table 1: Complete Function Word Categories (Leech et al. 2005)

Appendix B - Binary Ablation Results

Experiment Category	Feature Type	Ablation Method	Precision	Recall	F1	F1 Positive
Baseline			71.0%	67.4%	66.0%	73.0%
Feature Ablation	Function	Placeholder	69.5%	66.5%	65.1%	72.0%
Feature Ablation	Function	Remove	69.0%	65.9%	64.5%	71.6%
Feature Ablation	Hate	Placeholder	69.2%	66.1%	64.7%	71.8%
Feature Ablation	Hate	Remove	69.4%	66.3%	64.9%	71.9%
Random Words	Hate	Placeholder	69.6%	66.9%	65.7%	72.1%
Random Words	Hate	Remove	69.9%	67.5%	66.5%	72.3%
Feature Ablation	Neg_Sentiment	Placeholder	67.8%	65.4%	64.2%	70.7%
Feature Ablation	Neg_Sentiment	Remove	67.8%	65.4%	64.2%	70.8%
Random Words	Neg_Sentiment	Placeholder	71.0%	67.6%	66.2%	73.0%
Random Words	Neg_Sentiment	Remove	69.6%	66.7%	65.4%	72.1%
Feature Ablation	Pos_Sentiment	Placeholder	71.3%	68.1%	66.8%	73.3%
Feature Ablation	Pos_Sentiment	Remove	71.3%	68.2%	67.0%	73.3%
Random Words	Pos_Sentiment	Placeholder	71.5%	68.5%	67.4%	73.5%
Random Words	Pos_Sentiment	Remove	70.8%	67.7%	66.5%	72.9%

Table 2: MAMI Dataset: Binary Ablation Results (%)

Experiment Category	Feature Type	Ablation Method	Precision	Recall	F1	F1 Positive
Baseline			64.4%	63.9%	64.0%	69.7%
Feature Ablation	Function	Mask	62.6%	62.4%	62.4%	67.7%
Feature Ablation	Function	Remove	61.4%	61.3%	61.3%	66.3%
Feature Ablation	Hate	Mask	64.4%	63.7%	63.7%	70.3%
Feature Ablation	Hate	Remove	63.2%	62.6%	62.7%	69.0%
Random Words	Hate	Mask	62.5%	62.1%	62.1%	68.3%
Random Words	Hate	Remove	65.0%	64.9%	64.9%	69.4%
Feature Ablation	Neg_Sentiment	Mask	62.5%	62.1%	62.1%	68.3%
Feature Ablation	Neg_Sentiment	Remove	62.5%	62.1%	62.1%	68.3%
Random Words	Neg_Sentiment	Mask	64.4%	63.8%	63.9%	70.0%
Random Words	Neg_Sentiment	Remove	64.4%	63.9%	64.0%	69.7%
Feature Ablation	Pos_Sentiment	Mask	63.2%	62.6%	62.7%	69.0%
Feature Ablation	Pos_Sentiment	Remove	63.2%	62.5%	62.5%	69.3%
Random Words	Pos_Sentiment	Mask	63.2%	62.4%	62.3%	69.6%
Random Words	Pos_Sentiment	Remove	61.3%	60.9%	60.9%	67.3%

Table 3: EXIST2024 Dataset: Binary Ablation Results (%)

Appendix C - Fine-grained Multi-label Ablation Results

Experiment Category	Feature Type	Ablation Method	Shaming F1	Stereotype F1	Objectification F1	Violence F1	Macro average F1
Baseline			32.7%	44.8%	58.6%	42.8%	47.6%
Feature Ablation	All_Neg_Emotions	Placeholder	25.4%	43.1%	56.2%	35.0%	43.6%
Feature Ablation	All_Neg_Emotions	Remove	24.6%	44.0%	56.9%	36.3%	44.0%
Feature Ablation	Func_Auxiliary	Placeholder	33.7%	46.2%	57.3%	43.9%	48.3%
Feature Ablation	Func_Auxiliary	Remove	33.0%	45.4%	57.1%	41.9%	47.4%
Feature Ablation	Func_Conjunctions	Placeholder	33.6%	44.5%	56.5%	43.4%	47.3%
Feature Ablation	Func_Conjunctions	Remove	33.7%	44.5%	57.2%	44.9%	47.8%
Feature Ablation	Func_Determiners	Placeholder	31.6%	46.4%	57.4%	46.6%	48.1%
Feature Ablation	Func_Determiners	Remove	31.6%	46.2%	57.8%	46.2%	48.1%
Feature Ablation	Func_Enumerators	Placeholder	32.2%	45.3%	57.4%	43.2%	47.5%
Feature Ablation	Func_Enumerators	Remove	32.1%	46.1%	56.9%	43.9%	47.7%
Feature Ablation	Func_Interjections	Placeholder	32.6%	45.9%	58.0%	44.8%	48.1%
Feature Ablation	Func_Interjections	Remove	33.5%	45.9%	58.5%	44.0%	48.3%
Feature Ablation	Func_Particles	Placeholder	32.4%	45.3%	56.5%	43.6%	47.4%
Feature Ablation	Func_Particles	Remove	31.9%	45.3%	56.5%	44.3%	47.4%
Feature Ablation	Func_Prepositions	Placeholder	31.6%	45.9%	56.2%	45.2%	47.6%
Feature Ablation	Func_Prepositions	Remove	32.3%	45.7%	55.9%	45.2%	47.6%
Feature Ablation	Func_Pronouns	Placeholder	31.7%	44.4%	56.0%	45.0%	46.7%
Feature Ablation	Func_Pronouns	Remove	30.7%	44.1%	56.4%	45.7%	46.6%
Feature Ablation	Func_Qualifiers	Placeholder	32.7%	44.8%	57.2%	43.8%	47.4%
Feature Ablation	Func_Qualifiers	Remove	33.1%	45.5%	57.8%	44.4%	48.0%
Feature Ablation	Hate_AN	Placeholder	35.3%	44.6%	59.0%	43.8%	48.4%
Feature Ablation	Hate_AN	Remove	34.7%	44.0%	59.1%	43.5%	48.1%
Feature Ablation	Hate_ASF	Placeholder	32.7%	45.0%	57.3%	44.4%	47.6%
Feature Ablation	Hate_ASF	Remove	33.3%	45.7%	57.4%	44.0%	47.8%
Feature Ablation	Hate_ASM	Placeholder	33.0%	43.4%	58.3%	44.3%	47.6%
Feature Ablation	Hate_ASM	Remove	33.5%	43.9%	58.1%	43.8%	47.7%
Feature Ablation	Hate_CDS	Placeholder	34.3%	45.0%	58.1%	46.0%	48.8%
Feature Ablation	Hate_CDS	Remove	34.3%	44.5%	58.7%	44.3%	48.4%
Feature Ablation	Hate_DDF	Placeholder	30.5%	45.1%	58.1%	43.6%	47.2%
Feature Ablation	Hate_DDF	Remove	32.4%	44.7%	57.9%	43.8%	47.5%
Feature Ablation	Hate_DDP	Placeholder	32.2%	45.8%	57.4%	44.3%	47.9%
Feature Ablation	Hate_DDP	Remove	32.2%	45.6%	57.1%	44.3%	47.8%
Feature Ablation	Hate_DMC	Placeholder	31.7%	45.3%	57.5%	43.3%	47.4%
Feature Ablation	Hate_DMC	Remove	32.5%	44.4%	57.7%	43.3%	47.4%
Feature Ablation	Hate_IS	Placeholder	32.4%	45.5%	57.3%	44.9%	47.8%
Feature Ablation	Hate_IS	Remove	33.2%	45.6%	57.7%	44.3%	48.0%
Feature Ablation	Hate_OM	Placeholder	33.3%	44.2%	56.8%	44.0%	47.5%
Feature Ablation	Hate_OM	Remove	33.0%	45.5%	56.6%	42.9%	47.5%
Feature Ablation	Hate_OR	Placeholder	32.5%	45.1%	57.2%	45.1%	47.9%
Feature Ablation	Hate_OR	Remove	33.5%	45.4%	56.5%	43.6%	47.7%
Feature Ablation	Hate_PA	Placeholder	32.7%	45.6%	57.3%	44.3%	47.9%
Feature Ablation	Hate_PA	Remove	32.8%	45.2%	57.5%	44.3%	47.7%
Feature Ablation	Hate_PR	Placeholder	33.8%	43.8%	58.2%	42.2%	47.4%
Feature Ablation	Hate_PR	Remove	33.8%	43.5%	57.8%	42.1%	47.1%
Feature Ablation	Hate_PS	Placeholder	32.1%	45.6%	57.0%	45.0%	47.7%
Feature Ablation	Hate_PS	Remove	32.8%	45.4%	57.3%	46.4%	48.2%
Feature Ablation	Hate_QAS	Placeholder	32.7%	45.4%	57.4%	43.8%	47.7%
Feature Ablation	Hate_QAS	Remove	33.3%	45.1%	57.1%	43.8%	47.6%
Feature Ablation	Hate_RCI	Placeholder	32.7%	45.7%	57.8%	44.3%	48.0%
Feature Ablation	Hate_RCI	Remove	32.7%	45.6%	57.5%	44.3%	47.9%
Feature Ablation	Hate_RE	Placeholder	33.2%	45.3%	58.2%	39.8%	47.1%
Feature Ablation	Hate_RE	Remove	34.0%	46.0%	57.9%	39.1%	47.3%
Feature Ablation	Hate_SVP	Placeholder	33.2%	45.8%	57.6%	43.8%	47.9%
Feature Ablation	Hate_SVP	Remove	33.5%	45.4%	57.6%	44.3%	48.0%
Feature Ablation	Neg_Anger	Placeholder	31.2%	43.5%	58.0%	39.3%	46.0%
Feature Ablation	Neg_Anger	Remove	31.1%	44.2%	57.9%	36.9%	45.7%
Feature Ablation	Neg_Disgust	Placeholder	27.9%	44.2%	58.8%	39.6%	45.9%
Feature Ablation	Neg_Disgust	Remove	26.3%	44.0%	58.4%	37.9%	44.9%
Feature Ablation	Neg_Fear	Placeholder	31.7%	44.1%	57.9%	37.9%	45.9%
Feature Ablation	Neg_Fear	Remove	30.6%	43.8%	58.0%	36.0%	45.2%
Feature Ablation	Neg_Sadness	Placeholder	24.7%	43.5%	56.3%	37.3%	43.8%
Feature Ablation	Neg_Sadness	Remove	24.9%	43.7%	57.9%	34.9%	43.7%

Table 4: MAMI Dataset: Subcategory and Macro F1 Scores (%)

Experiment Category	Feature Type	Ablation Method	Ideological-inequality F1	Stereotyping-dominance F1	Objectification F1	Sexual-violence F1	Misogyny-non-sexual-violence F1	Macro average F1
Baseline			53.7%	29.9%	50.0%	33.3%	17.4%	40.4%
Feature Ablation	All_Neg_Emotions	Mask	53.7%	31.1%	44.7%	26.1%	8.7%	37.2%
Feature Ablation	All_Neg_Emotions	Remove	51.9%	31.5%	50.0%	26.1%	8.3%	37.9%
Feature Ablation	Func_Auxiliary	Mask	53.0%	34.0%	47.4%	32.0%	17.4%	40.4%
Feature Ablation	Func_Auxiliary	Remove	54.3%	34.0%	44.7%	24.0%	17.4%	39.0%
Feature Ablation	Func_Conjunctions	Mask	52.5%	29.2%	52.7%	33.3%	8.7%	39.1%
Feature Ablation	Func_Conjunctions	Remove	51.2%	30.4%	52.2%	33.3%	16.7%	40.0%
Feature Ablation	Func_Determiners	Mask	51.2%	31.5%	52.2%	33.3%	27.3%	42.3%
Feature Ablation	Func_Determiners	Remove	50.6%	29.5%	53.3%	34.8%	28.6%	42.6%
Feature Ablation	Func_Enumerators	Mask	53.7%	29.8%	47.8%	33.3%	17.4%	39.8%
Feature Ablation	Func_Enumerators	Remove	53.7%	30.1%	50.0%	33.3%	17.4%	40.2%
Feature Ablation	Func_Interjections	Mask	53.0%	30.4%	51.6%	25.0%	16.7%	38.8%
Feature Ablation	Func_Interjections	Remove	53.0%	30.4%	50.0%	25.0%	16.7%	38.2%
Feature Ablation	Func_Particles	Mask	53.7%	32.6%	50.0%	30.8%	8.7%	38.5%
Feature Ablation	Func_Particles	Remove	53.7%	28.9%	50.5%	30.8%	16.7%	39.3%
Feature Ablation	Func_Prepositions	Mask	53.0%	32.3%	50.5%	30.8%	24.0%	40.7%
Feature Ablation	Func_Prepositions	Remove	52.4%	32.6%	48.9%	23.1%	24.0%	39.3%
Feature Ablation	Func_Pronouns	Mask	55.4%	31.9%	51.8%	40.0%	19.0%	42.7%
Feature Ablation	Func_Pronouns	Remove	54.1%	30.3%	46.7%	41.7%	18.2%	40.7%
Feature Ablation	Func_Qualifiers	Mask	53.7%	30.4%	51.6%	38.5%	17.4%	41.4%
Feature Ablation	Func_Qualifiers	Remove	53.7%	30.8%	49.5%	38.5%	17.4%	41.2%
Feature Ablation	Hate_AN	Mask	53.7%	31.8%	50.5%	32.0%	24.0%	41.8%
Feature Ablation	Hate_AN	Remove	53.0%	33.0%	50.5%	32.0%	24.0%	41.6%
Feature Ablation	Hate_ASF	Mask	53.7%	32.3%	50.0%	33.3%	17.4%	40.6%
Feature Ablation	Hate_ASF	Remove	53.7%	32.3%	48.9%	33.3%	16.7%	40.2%
Feature Ablation	Hate_ASM	Mask	52.4%	34.8%	53.6%	32.0%	24.0%	42.7%
Feature Ablation	Hate_ASM	Remove	51.8%	36.2%	51.5%	32.0%	17.4%	41.0%
Feature Ablation	Hate_CDS	Mask	52.4%	32.6%	49.5%	25.0%	14.8%	38.3%
Feature Ablation	Hate_CDS	Remove	53.7%	32.6%	50.5%	25.0%	15.4%	38.7%
Feature Ablation	Hate_DDF	Mask	53.7%	28.6%	51.6%	32.0%	16.7%	40.0%
Feature Ablation	Hate_DDF	Remove	53.7%	30.1%	52.2%	32.0%	16.7%	40.3%
Feature Ablation	Hate_DDP	Mask	53.7%	33.0%	51.1%	30.8%	16.7%	40.3%
Feature Ablation	Hate_DDP	Remove	53.7%	32.6%	51.1%	30.8%	16.7%	40.3%
Feature Ablation	Hate_DMC	Mask	53.7%	31.1%	50.5%	32.0%	16.0%	39.9%
Feature Ablation	Hate_DMC	Remove	53.0%	30.4%	52.2%	32.0%	15.4%	39.9%
Feature Ablation	Hate_IS	Mask	53.7%	30.8%	52.7%	32.0%	16.7%	40.5%
Feature Ablation	Hate_IS	Remove	53.7%	28.6%	51.6%	32.0%	17.4%	39.9%
Feature Ablation	Hate_OM	Mask	53.7%	29.5%	51.6%	32.0%	9.1%	38.9%
Feature Ablation	Hate_OM	Remove	53.0%	27.7%	51.1%	32.0%	16.0%	39.3%
Feature Ablation	Hate_OR	Mask	53.7%	30.1%	50.0%	33.3%	16.7%	40.0%
Feature Ablation	Hate_OR	Remove	53.7%	30.1%	50.0%	33.3%	17.4%	40.1%
Feature Ablation	Hate_PA	Mask	54.3%	30.1%	50.5%	32.0%	16.7%	40.3%
Feature Ablation	Hate_PA	Remove	54.3%	30.1%	50.5%	32.0%	16.7%	40.3%
Feature Ablation	Hate_PR	Mask	52.4%	32.6%	51.6%	30.8%	16.0%	40.3%
Feature Ablation	Hate_PR	Remove	53.7%	30.8%	50.5%	32.0%	16.0%	39.9%
Feature Ablation	Hate_PS	Mask	53.7%	30.4%	51.6%	32.0%	16.7%	40.1%
Feature Ablation	Hate_PS	Remove	53.7%	30.4%	51.1%	32.0%	16.7%	40.0%
Feature Ablation	Hate_QAS	Mask	53.7%	28.6%	50.5%	32.0%	16.7%	39.8%
Feature Ablation	Hate_QAS	Remove	53.0%	33.0%	50.5%	32.0%	16.7%	40.4%
Feature Ablation	Hate_RCI	Mask	53.7%	30.4%	50.5%	32.0%	16.7%	40.1%
Feature Ablation	Hate_RCI	Remove	53.7%	30.1%	50.5%	32.0%	16.7%	40.1%
Feature Ablation	Hate_RE	Mask	53.7%	32.3%	48.9%	26.1%	17.4%	39.1%
Feature Ablation	Hate_RE	Remove	54.3%	32.3%	49.5%	26.1%	16.7%	39.4%
Feature Ablation	Hate_SVP	Mask	51.9%	30.8%	51.1%	32.0%	16.7%	40.1%
Feature Ablation	Hate_SVP	Remove	53.7%	30.4%	54.3%	30.8%	16.7%	40.6%
Feature Ablation	Neg_Anger	Mask	51.2%	32.6%	45.7%	26.1%	8.0%	36.7%
Feature Ablation	Neg_Anger	Remove	51.2%	32.6%	47.3%	27.3%	8.0%	37.3%
Feature Ablation	Neg_Disgust	Mask	50.0%	31.1%	47.3%	32.0%	16.7%	38.8%
Feature Ablation	Neg_Disgust	Remove	50.6%	30.8%	47.8%	32.0%	17.4%	39.2%
Feature Ablation	Neg_Fear	Mask	53.7%	29.5%	52.3%	32.0%	16.7%	40.3%
Feature Ablation	Neg_Fear	Remove	52.5%	29.2%	52.7%	32.0%	16.7%	40.2%
Feature Ablation	Neg_Sadness	Mask	55.4%	28.3%	50.0%	33.3%	17.4%	40.1%
Feature Ablation	Neg_Sadness	Remove	53.7%	28.3%	50.5%	33.3%	17.4%	39.9%

Table 5: EXIST2024 Dataset: Subcategory and Macro F1 Scores (%)

References

W. Alorainy, P. Burnap, H. Liu, and M. L. Williams. "the enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web*, 13(3), July 2019. ISSN 1559-1131. doi:10.1145/3324997. URL https://doi.org/10.1145/3324997.

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. *CoRR*, abs/1706.00188, 2017. URL http://arxiv.org/abs/1706.00188.

E. Bassignana, V. Basile, and V. Patti. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It 2018)*. CEUR Workshop Proceedings, 2018.

R. S. Bigler and C. Leaper. Gendered language: Psychological principles, evolving practices, and inclusive policies. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):187–194, 2015. doi:10.1177/2372732215600452. URL https://doi.org/10.1177/2372732215600452.

T. Caselli, V. Basile, J. Mitrović, and M. Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, Aug. 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.woah-1.3. URL https://aclanthology.org/2021.woah-1.3/.

P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti. Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, 14(1):322–352, 2022. doi:10.1007/s12559-021-09862-5. URL https://doi.org/10.1007/s12559-021-09862-5.

T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017. URL http://arxiv.org/abs/1703.04009.

M.-C. de Marneffe, C. D. Manning, J. Nivre, and D. Zeman. Universal dependencies. *Computational Linguistics*, 47(2):255–308, 2021. doi:10.1162/coli_a_00402. URL https://doi.org/10.1162/coli_a_00402.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, and J. Sorensen. SemEval-2022 task 5: Multimedia automatic misogyny identification. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.semeval-1.74. URL https://aclanthology.org/2022.semeval-1.74/.

S. Hakimov, G. S. Cheema, and R. Ewerth. TIB-VA at SemEval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 756–760, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.semeval-1.105. URL https://aclanthology.org/2022.semeval-1.105/.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *CoRR*, abs/2005.04790, 2020. URL https://arxiv.org/abs/2005.04790.

G. Leech, M. Deuchar, and R. Hoogenraad. *English Grammar for Today: A New Introduction*. Palgrave Macmillan, 2nd edition, 2005. Originally published in 1982.

D. A. Lestari, M. Primagara, S. A. Sari, A. Meilina, S. Fauziah, A. I. Sugesti, A. Nasywa, and A. D. Salwi. Meme culture: A study of humor and satire in digital media. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(4):134–140, 2024. doi:10.62225/2583049X.2024.4.4.3013. URL https://doi.org/10.62225/2583049X.2024.4.4.3013.

J. Li, D. Li, C. Xiong, and S. C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. URL https://arxiv.org/abs/2201.12086.

J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

H. Lin, Z. Luo, W. Gao, J. Ma, B. Wang, and R. Yang. Towards explainable harmful meme detection through multimodal debate between large language models, 2024. URL https://arxiv.org/abs/2401.13298.

I. Markov, N. Ljubešić, D. Fišer, and W. Daelemans. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In O. De Clercq, A. Balahur, J. Sedoc, V. Barriere, S. Tafreshi, S. Buechel, and V. Hoste,

editors, *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online, Apr. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wassa-1.16/.

S. M. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297, 2013. URL http://arxiv.org/abs/1308.6297.

F. Naznin, M. T. Rahman, and S. R. Alve. Hierarchical sentiment analysis framework for hate speech detection: Implementing binary and multiclass classification strategy, 2024. URL https://arxiv.org/abs/2411.05819.

D. Njagi, Z. Zuping, D. Hanyurwimfura, and J. Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015. doi:10.14257/ijmue.2015.10.4.21. URL https://www.researchgate.net/publication/283125668_A_Lexicon-based_Approach_for_Hate_Speech_Detection.

J. W. Pennebaker. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press, 2011. doi:10.1016/S0262-4079(11)62167-2. URL https://doi.org/10.1016/S0262-4079(11)62167-2.

L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, and D. Spina. Overview of exist 2024 – learning with disagreement for sexism identification and characterization in tweets and memes. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 93–117, Grenoble, France, September 2024. Springer Nature Switzerland. ISBN 978-3-031-71908-0. doi:10.1007/978-3-031-71908-0_5.

F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña López, and M. T. Martín-Valdivia. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Trans. Internet Technol.*, 20(2), Mar. 2020. ISSN 1533-5399. doi:10.1145/3369869. URL https://doi.org/10.1145/3369869.

S. Pramanick, D. Dimitrov, R. Mukherjee, S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty. Detecting harmful memes and their targets. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online, Aug. 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.findings-acl.246. URL https://aclanthology.org/2021.findings-acl.246/.

A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327, 2020. URL https://arxiv.org/abs/2002.12327.

A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing. In L.-W. Ku and C.-T. Li, editors, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. doi:10.18653/v1/W17-1101. URL https://aclanthology.org/W17-1101/.

M. Sharma, I. Kandasamy, and V. W B. R2D2 at SemEval-2022 task 5: Attention is only as good as its values! a multimodal system for identifying misogynist memes. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 761–770, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.semeval-1.106. URL https://aclanthology.org/2022.semeval-1.106/.

L. Shifman. *Memes in Digital Culture*. MIT Press Essential Knowledge. MIT Press, Cambridge, MA, 2014. ISBN 9780262525435.

Universal Dependencies Consortium. Universal pos tags, 2024. URL https://universaldependencies.org/u/pos/all.html. Accessed: 2025-06-23.

J. Van Nooten, I. Markov, and W. Daelemans. Evaluating the impact of word classes on cross-domain age detection models' performance. *Computational Linguistics in the Netherlands Journal*, 11:71–84, 2021. URL https://clinjournal.org/clinj/article/view/122.

B. E. Wiggins. *The Discursive Power of Memes in Digital Culture: Ideology, Semiotics, and Intertextuality*. Routledge, 2019. ISBN 9780429492303. doi:10.4324/9780429492303.

M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117, January 2020. doi:10.1093/bjc/azz049. URL https://doi.org/10.1093/bjc/azz049.

J. Zhang and Y. Wang. SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.semeval-1.81. URL https://aclanthology.org/2022.semeval-1.81/.

J. Zhi, Z. Mengyuan, M. Yuan, D. Hu, X. Du, L. Jiang, Y. Mo, and X. Shi. PAIC at SemEval-2022 task 5: Multi-modal misogynous detection in MEMES with multi-task learning and multi-model fusion. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, and S. Ratan, editors, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 555–562, Seattle, United States, July 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.semeval-1.76. URL https://aclanthology.org/2022.semeval-1.76/.