Research Master Thesis

Exploring Ensemble Strategies for Misogynous and Sexist Meme Detection

Ariana Britez

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics

(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by:	dr. Ilia Markov
2^{nd} reader:	dr. Hennie van der Vliet
Submitted:	August 11, 2025

Abstract

Due to the wide spread of sexism and misogyny in online platforms in the form of memes, this thesis focused on the identification and classification of sexism and misogyny in multimodal content. Since previous research showed that ensembles of conventional machine learning and deep learning models have proven useful for the classification of textual harmful content, they were explored to determine their benefits in the identification and categorization of sexism and misogyny in memes.

This was explored through a hard majority voting ensemble strategy that combined conventional machine learning and deep learning models using the datasets that originated in two shared tasks on sexism and misogyny. The tasks involved binary classification in which memes were classified as sexist or misogynous or not, depending on the dataset used, and a hierarchical multi-label classification in which the fine-grained categories of sexism and misogyny were identified. Evaluation was conducted in the typical in-domain setup as well as in a cross-dataset setup to asses generalization across related datasets. The models combined in the ensemble strategy included two models based on the textual modality and a state-of-the-art multimodal model. The text-only models, which processed the meme text and image captions as inputs, were an SVM with stylometric and emotion-based features, and a fine-tuned RoBERTa model. The meme image and meme text were processed with a multimodal model that combined Swin Transformer V2 with RoBERTa followed by an MLP fusion model and prediction layer for classification. The multimodal approach also served as as baseline to quantify the improvements brought by the ensemble strategy.

The results revealed that an ensemble strategy of conventional machine learning and deep learning models showed the best performance in multimodal binary sexism identification in in-domain and cross-dataset setups. While the ensemble did not outperform the component models in multi-label sexism categorization or in binary and multi-label misogyny classification across in-domain and cross-dataset settings, it achieved results comparable to those of the individual models. An error analysis revealed the most common patterns of misclassification by the ensemble. The use of rhetorical devices such as sarcasm and irony accounted for the highest percentage of false negatives across both datasets in in-domain classification. In cross-dataset misogyny identification, however, incorrect or poor suggestive image descriptions were the primary cause of false negatives. Regarding false positives, the most frequent cause across both in-domain and cross-dataset settings was the presence of women in memes, either in the image or in the text.

Declaration of Authorship

I, Ariana Judith Britez, declare that this thesis, titled *Exploring Ensemble Strategies for Misogynous and Sexist Meme Detection* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree 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: August 11, 2025

Signed:

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, dr. Ilia Markov, for his guidance, support, and patience throughout the entire process of researching and writing this thesis. From the moment I proposed the idea, he helped shape it and provided valuable feedback that pushed me to take my work further than I thought possible. This thesis would not have been possible without his trust and assistance.

Thanks for dr. Elisabetta Fersini for suggesting the EXIST 2024 meme dataset for the experiments I conducted in this thesis.

I am also grateful to the members of the CLTL Lab at VU Amsterdam for their guidance through my master's programme and for all that I have learned about the field of computational linguistics over the past two years.

Many thanks to my fellow students Farnaz Bani Fatemi, Christina Karavida, Szabolcs Pál, Ruth Scheichl, and Hannah Heath for all our teamwork and for making this master's programme more enjoyable.

Finally, I am incredibly grateful to my husband, Sjef, whose unwavering support and encouragement sustained me throughout this journey and made all the difference. To my mother, thank you for always believing in me. I also wish to express my appreciation to my friends for cheering me on throughout this process.

List of Figures

1.1 Examples of misogynous and non-misogynous memes (taken from Rehman et al. (2025)). . . . 3

3.1 Distribution of binary (top) and fine-grained classes when meme is misogynous (bottom) in the MAMI training and test sets. . . . 19
3.2 Distribution of binary (top) and fine-grained classes when meme is sexist (bottom) in the EXIST 2024 dataset. . . . 22
3.3 MAMI dataset splits after stratified sampling for in-domain experiments. Distribution of binary (top) and fine-grained classes when meme is misogynous (bottom). . . . 23
3.4 EXIST 2024 dataset splits after stratified sampling for in-domain experiments. Distribution of binary (top) and fine-grained classes when meme is sexist (bottom). . . . 24
3.5 EXIST 2024 dataset splits after stratified sampling for cross-dataset experiments. Distribution of binary (top) and fine-grained classes when meme is sexist (bottom). . . . 26
3.6 MAMI dataset splits after stratified sampling for cross-dataset experimentss. Distribution of binary (top) and fine-grained classes when meme is misogynous (bottom). . . . 27
3.7 ViT architecture (taken from Dosovitskiy et al. (2021)) . . . 36
3.8 A comparison of Swin and ViT feature mapping (taken from Liu et al. (2021)) . . . 37
3.9 Overall architecture of implemented ensemble approach. The multimodal architecture (upper part of the figure) is based on Wang and Markov (2024c). . . . 38

4.1 Confusion matrices for the component models and ensemble approach in binary classification on EXIST 2024. . . . 43
4.2 Confusion matrices for the component models and ensemble approach in binary classification on MAMI. . . . 43
4.3 Confusion matrices for the component models and ensemble approach in binary classification when training on MAMI and testing on EXIST 2024. 44
4.4 Confusion matrices for the component models and ensemble approach in binary classification when training on EXIST 2024 and testing on MAMI. 44
4.5 Confusion matrices for the component models and ensemble approach in multi-label classification on EXIST 2024. avg: Average of predicted fine-grained classes. . . . 46

4.6 Confusion matrices for the component models and ensemble approach in multi-label classification on MAMI. avg: Average of predicted fine-grained classes. . . . 46
4.7 Confusion matrices for the component models and ensemble approach in multi-label classification when training on MAMI and testing on EXIST 2024. avg: Average of predicted fine-grained classes. . . . 48
4.8 Confusion matrices for the component models and ensemble approach in multi-label classification when training on EXIST 2024 and testing on MAMI. avg: Average of predicted fine-grained classes. . . . 48
4.9 Examples of FN errors of the type sexism or misogyny without swear or offensive words. . . . 51
4.10 Examples of FN errors of the type rhetorical questions. . . . 52
4.11 Examples of FN errors of the type metaphors and comparisons. . . . 53
4.12 Examples of FN errors of the type sarcasm and irony. . . . 53
4.13 Examples of FN errors of the type suggestive or explicit image. . . . 54
4.14 Examples of FN errors of the type quality of the meme text. . . . 55
4.15 Examples of FP errors of the type doubtful labels. . . . 56
4.16 Examples of FP errors of the type swear or offensive words. . . . 56
4.17 Examples of FP errors of the type women in meme. . . . 57
4.18 Examples of FP errors of the type men in meme. . . . 58
4.19 Examples of FP errors of the type domestic chores. . . . 58
4.20 Examples of FP errors of the type reference to violence. . . . 59

B.1 Confusion matrices per fine-grained category per model implemented on the EXIST 2024 dataset. Non-sexist category predicted with Ensemble in the first step. . . . 69
B.2 Confusion matrices per fine-grained category per model implemented on the MAMI dataset. Non-misogynous category predicted with Ensemble in the first step. . . . 70
B.3 Confusion matrices per fine-grained category per model when training on MAMI dataset and testing on EXIST 2024 dataset. Non-sexist category predicted with Ensemble in the first step. . . . 71
B.4 Confusion matrices per fine-grained category per model when training on EXIST 2024 dataset and testing on MAMI dataset. Non-misogynous category predicted with Ensemble in the first step. . . . 72

List of Tables

3.1 Statistics of the MAMI training and test sets for misogyny identification. 18
3.2 Statistics of the MAMI training and test sets for misogyny categorization. Percentages considered in terms of the binary label. 19
3.3 Statistics of the EXIST 2024 training set for sexism identification. # with Cat: Number of memes with fine-grained classes. 20
3.4 Statistics of the EXIST 2024 training set for sexism categorization. Percentages considered in terms of the *sexist* label. 21
3.5 MAMI binary class distribution after stratified sampling for in-domain experiments. *Test from original dataset. 23
3.6 MAMI fine-grained class distribution after stratified sampling for in-domain experiments. *Test from original dataset. 23
3.7 EXIST 2024 binary class distribution after stratified sampling for in-domain experiments. 24
3.8 EXIST 2024 fine-grained class distribution after stratified sampling for in-domain experiments. 24
3.9 EXIST 2024 binary class distribution after stratified sampling for cross-dataset experiments. $\mathrm{T_o}$: Total with overlapping classes. 25
3.10 EXIST 2024 fine-grained class distribution after stratified sampling for cross-dataset experiments. $\mathrm{T_o}$: Total with overlapping classes. 25
3.11 MAMI binary class distribution after stratified sampling for cross-dataset experiments. $\mathrm{T_o}$: Total with overlapping classes. 27
3.12 MAMI fine-grained class distribution after stratified sampling for cross-dataset experiments. $\mathrm{T_o}$: Total with overlapping classes. 27
3.13 Meme, meme text and image caption from the BLIP-2 model. 28
3.14 Performance of different parameter configurations across features implemented in SVM approach on EXIST 2024 dataset. POS: part-of-speech; FW: function words; Emo: emotion conveying words; WS: weighting scheme. The selected configuration for the SVM approach is shown in pink. 31
3.15 Performance of explored pre-trained models for sexism identification on EXIST 2024 dataset. 34
3.16 Best parameters found with Optuna to fine-tune RoBERTa for each task. 35
3.17 Performance of the multimodal models explored for sexism identification on EXIST 2024 dataset. 37

4.1 In-domain and cross-dataset results for the multimodal baseline, component models, and ensemble in binary classification. Pos. Class: F1 score of positive class on each dataset. 42

4.2 In-domain and cross-dataset results for the multimodal baseline, component models, and ensemble in multi-label classification. WA: Weighted-averaged F1 score. 45

4.3 Correlations between predictions of component models with Pearson correlation coefficient. The values in multi-label classification are averaged across classes, and those with $r < 0.5$ are highlighted in bold type. . . . 50

A.1 Total phrases removed from image captions across datasets. 67

Contents

Abstract i

Declaration of Authorship iii

Acknowledgments v

List of Figures vii

List of Tables ix

1 Introduction 1

1.1 Research Problem . . . 2

1.2 Research Questions . . . 4

1.3 Main Contributions . . . 5

1.4 Thesis Structure . . . 5

2 Related Work 7

2.1 Detecting Harmful Content in Textual Data . . . 7

2.1.1 Conventional Machine Learning Approaches . . . 7

2.1.2 Transformer Models . . . 10

2.1.3 Ensemble Strategies . . . 10

2.2 Detecting Harmful Content in Multimodal Data . . . 12

2.2.1 Approaches for Detecting Harmful Memes . . . 12

2.2.2 Ensemble Strategies for Multimodal Harmful Content Detection 14

2.3 Research Focus . . . 15

3 Methodology 17

3.1 Datasets . . . 17

3.1.1 MAMI . . . 17

3.1.2 EXIST 2024 . . . 20

3.1.3 Stratified Sampling . . . 21

3.2 Preprocessing . . . 28

3.3 Experimental Setup & Models . . . 29

3.3.1 Text-only Approaches . . . 29

3.3.2 Multimodal Approach . . . 35

3.3.3 Ensemble Strategy . . . 38

3.4 Evaluation Metrics . . . 39

4 Results and Analyses 41

4.1 Quantitative Analysis . . . 41

4.1.1 Binary Classification . . . 41

4.1.2 Multi-label Classification . . . 45

4.2 Qualitative Analysis . . . 49

4.2.1 Correlation Analysis . . . 49

4.2.2 Error Analysis . . . 50

5 Conclusion and Discussion 61

5.1 Conclusions . . . 61

5.1.1 Addressing Research Questions . . . 62

5.2 Limitations . . . 64

5.3 Future Work . . . 65

A Preprocessing: Image Captioning 67

A.1 Phrases Removed in Image Captions . . . 67

B Multi-Label Classification: Confusion Matrices 69

B.1 In-Domain . . . 69

B.1.1 EXIST 2024 . . . 69

B.1.2 MAMI . . . 70

B.2 Cross-Dataset . . . 71

B.2.1 EXIST 2024 . . . 71

B.2.2 MAMI . . . 72

Chapter 1

Introduction

Online platforms such as social media and discussion forums provide users with tools to create and share a wide range of information. However, the rise of user-generated content has also been accompanied by an increase in hateful content. Focusing on gender, the Pew Research Center (Duggan, 2017) reported that women were twice as likely as men to have experienced gender-based online harassment, which can manifest in the form of *sexist* hate speech or *misogyny*. As explained by Cameron (2023), *sexism* refers to prejudice or discrimination against anyone on the basis of their sex, a definition that suggests that men can also be victims of sexism. While some dictionaries state that sexism "typically" or "especially" targets women, feminists argue that such a definition fails to capture the political context of sexism as a systemic issue against women and specify that sexism is "directed against women". The definition of *misogyny*, on the other hand, is explicitly sex-specific and refers to hatred of women. However, some dictionaries also define it as prejudice against women, which overlaps with the meaning of *sexism* and suggests that the term has acquired a weaker sense over time. Both of these phenomena work hand-in-hand to uphold patriarchal social structures.[1] Rodríguez-Sánchez et al. (2020) differentiate them with a clear example: when a man claims that he prefers his wife to stay at home because she can take better care of their children, he is being sexist. In contrast, when a man claims that women should only be allowed to stay at home, he is being misogynist.

Both sexism and misogyny are expressed not only in comments or posts but also in multimodal memes. While memes, which combine images with text added *a posteriori*, are often used for humorous or ironic effects, they are also employed to spread violence and aggression against women (Paciello et al., 2021). This form of gender-based violence is considered an abuse to human rights,[2] and the effects of sexist and misogynous content can impact not only targeted victims but also non-targeted groups and society as a whole (Cervone et al., 2021). Regarding targeted victims, the spread of misogyny and sexism can affect them on professional, psychological, and personal levels. Professionally, it can force individuals to leave online spaces where they conduct business or even abandon their careers. Psychologically, it can lead to fear, anxiety, eating disorders, and even suicidal thoughts and attempts. On a personal level, it often results in reduced online presence, limiting participation and visibility, and contributing to social isolation through the loss of both personal and professional interactions

[1] https://www.vox.com/identities/2017/12/5/16705284/elizabeth-warren-loss-2020-sexism-misogyny-kate-manne

[2] https://www.amnesty.org/en/what-we-do/technology/online-violence/

(Poland, 2016). As for non-targeted groups and society, on the other hand, increased exposure to this type of content can lead to desensitization, as it appears less offensive and more socially acceptable to them. Those who are frequently exposed to it–in particular young people–might perceive other non-normative behaviours as more socially and morally acceptable, and even worthy of imitation (Cervone et al., 2021).

After social media platforms such as X and Meta reduced their moderation efforts, allowing the spread of sexism and misogyny, and other forms of hateful content,[3] there has been an increase in such content online.[4] However, users expect these platforms to take an active role in limiting harmful content and misinformation, considering moderation essential to maintain healthy digital spaces (Theocharis et al., 2025). Due to the rise in harmful content and the serious impact that sexism and misogyny can have on victims, the automatic detection of sexist and misogynous memes becomes increasingly important. Their spread could not only amplify social misbehaviour by supporting and even inciting hate crimes (Frenda et al., 2019), but also contribute to sexual stereotyping and gender inequalities offline (Fersini et al., 2021). Furthermore, identifying the different forms in which sexism and misogyny occur in memes could help recognize patterns in how they manifest online (Plaza et al., 2024a). Their automatic detection not only aims to contribute to reducing harmful content online but also to promote a fair treatment of women both in online and offline spaces.

1.1 Research Problem

Since memes consist of two different atomic units of information, i.e. image and text, they represent a multimodal classification problem. Memes are particularly challenging in this context because the same image can be paired with different text, or the same text can appear with different images, potentially conveying harmful or non-harmful content in each case. Moreover, toxicity or hate can be conveyed through the image alone, the text alone, or a combination of both (Fersini et al., 2021). This complexity is illustrated in Figure 1.1, where two memes using the same image but different text convey (a) a misogynistic instance and (b) a non-misogynistic one; this is also the case in (c) and (d), where the same text is combined with different images.

Given that previous research demonstrated that ensemble strategies of shallow and deep learning approaches improve performance in hate speech detection for textual content (van Aken et al., 2018; Markov et al., 2021; Markov and Daelemans, 2021; Markov et al., 2022), this thesis aims to explore whether similar benefits can be achieved in multimodal scenarios. Therefore, the proposed approach involves an ensemble that combines multimodal and text-only models to address the identification and categorization of sexism and misogyny in memes. The datasets used originated from two shared tasks: MAMI (Multimedia Automatic Misogyny Identification) from SemEval-2022 Task 5 (Fersini et al., 2022) and EXIST (sEXism Identification in Social neTworks) from CLEF 2024 (Plaza et al., 2024a,b). The first task was a binary classification task, where a meme was classified as either sexist or non-sexist in the EXIST 2024 dataset, and as misogynous or non-misogynous in the MAMI dataset. The second task was a multi-label classification task, in which each meme was assigned one or more overlapping fine-grained categories of sexism or misogyny.

[3] https://www.amnesty.org/en/latest/news/2025/02/meta-new-policy-changes/

[4] https://news.berkeley.edu/2025/02/13/study-finds-persistent-spike-in-hate-speech-on-x/, https://transparency.meta.com/en-gb/integrity-reports-q1-2025/

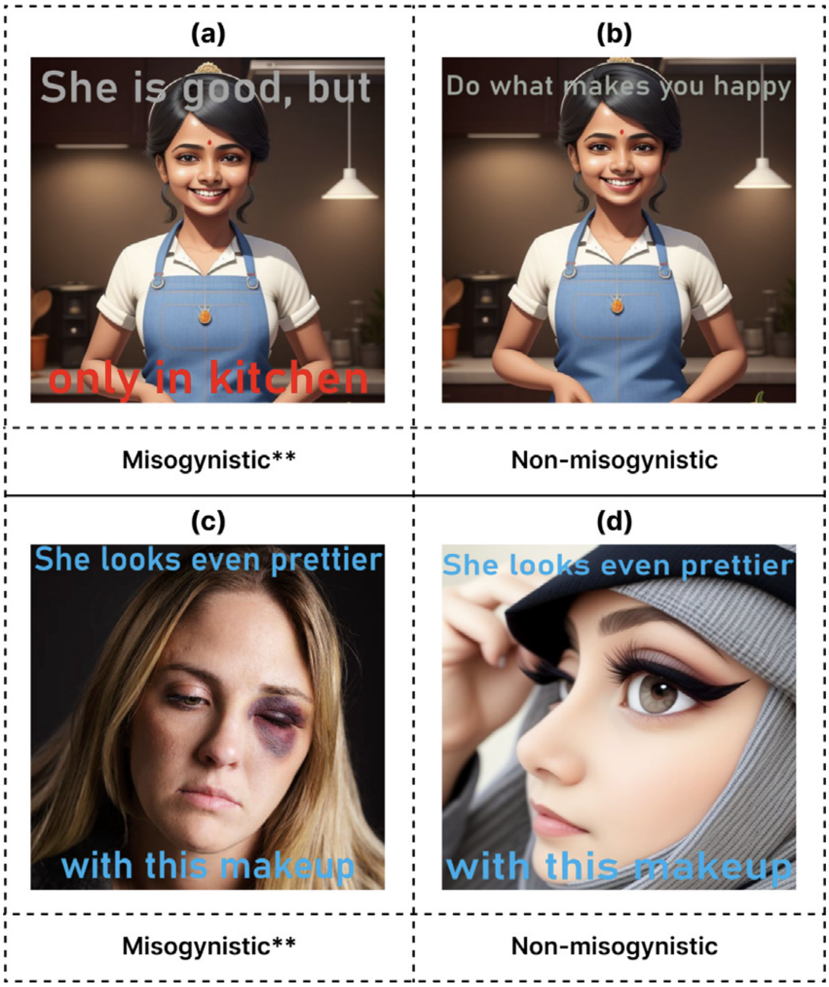


Figure 1.1: Examples of misogynous and non-misogynous memes (taken from Rehman et al. (2025)).

Apart from evaluating the performance on each dataset in the typical in-domain setup, a key focus of this study is the cross-dataset evaluation to assess how well the models generalize to another, yet related, task. Unlike using a single dataset for training and evaluation, which might not reflect real-word scenarios, cross-dataset evaluation assess the ability of a model to generalize beyond its training data (Ramis et al., 2022). By revealing whether the model has learned transferable patterns or has overfit to the specific characteristics of one dataset, this evaluation approach provides a more realistic assessment of its robustness in real-word applications. When working with each dataset in the in-domain setup, the fine-grained categories specific to each dataset were preserved. However, for the cross-dataset evaluation setup, only the overlapping classes between the two datasets were considered. This evaluation setup has not been previously applied to meme datasets related to misogyny and sexism, particularly in the multi-label classification task, as previous cross-domain experiment on the MAMI dataset focused only on the binary classification task (Aggarwal et al., 2024).

1.2 Research Questions

The objective of this thesis is to address the automatic detection of misogynous and sexist memes through the following research question:

RQ: Given that ensemble strategies have shown promising results for detecting harmful content in textual data, including in cross-domain setups, are they also beneficial for the detection and classification of sexism and misogyny in multimodal data?

In order to answer this main question, the following sub-questions are posed:

1. What are the best component models to be incorporated into the ensemble?

2. Is the ensemble strategy helpful for both binary and multi-label classification across different datasets?

3. Is the ensemble strategy useful for the in-domain setup as well as cross-dataset setting?

4. Given that shallow approaches like SVM help reduce the false positive rate in text, do they also contribute to a lower false positive rate in multimodal data when incorporated into an ensemble?

5. Based on error and correlation analyses, why are ensemble strategies helpful?

It is hypothesized that models based on the textual modality–which includes the meme text and an image caption that provides a textual representation of the meme image–combined with a multimodal model will be the most effective component models to include in an ensemble. In particular, the text-only models considered will be an SVM model that incorporates stylometric and emotion-based features (Markov et al., 2021) and an encoder-only model. The multimodal model will combine the image and text embeddings from the memes (Wang and Markov, 2024a). The performance of the ensemble is expected to surpass the current state-of-the-art in multimodal harmful content detection. Moreover, it is hypothesized that ensemble strategies will improve performance in both in-domain and cross-dataset evaluation setups, for both binary and multi-label classification tasks. Regarding the false positive rate associated with the SVM model, it is anticipated that a similar pattern will be observed when classifying memes.

Considering that both textual and multimodal content often follow similar patterns in expressing harmful messages, such as the use of swear words, insults, threats, profane language (Zampieri et al., 2019), or emotionally charged expressions (Markov et al., 2021), it is expected that methods that proved effective in detecting textual harmful content will also generalize to multimodal scenarios. Moreover, since the source of sexism or misogyny in memes can lie in the text, the image, or their combination (Fersini et al., 2021), integrating methods that capture both textual and visual cues can enhance performance. Therefore, employing approaches that effectively handle textual harmful content–using the meme text and image captions as textual representations–combined with a multimodal approach that processes the visual and textual modalities might offer a comprehensive and effective strategy to address the identification and classification of sexism and misogyny in multimodal content.

To test the above-mentioned hypotheses, various component models, including text-based and multimodal models, will be trained (in the case of the conventional machine learning approach) or fine-tuned (in the case of transformer-based models) and combined into a hard majority voting ensemble strategy to address the identification and categorization of sexism and misogyny in memes. In a hard majority voting ensemble, the discrete predictions of multiple models are combined and the final output is determined by the label receiving the most votes, based on the premise that the errors of individual models can be compensated by the others (Sagi and Rokach, 2018). The ensembles will also be evaluated in a cross-dataset setup, which is crucial for assessing the generalizability and robustness of a model beyond its training data, thereby reflecting real-world applications. A correlation analysis together with an error analysis will provide further insight into why ensemble strategies are beneficial for the detection and classification of sexism and misogyny in memes.

1.3 Main Contributions

The main contributions of this work could be summarized as follows:

1. To develop an approach for the identification and categorization of sexism and misogyny in multimodal content using an ensemble of deep learning and conventional machine learning models, including an SVM based on stylometric and emotion-based features;

2. to investigate the performance of an ensemble in in-domain and cross-dataset setups in both binary and multi-label classification;

3. to examine the reasons as to why the ensemble strategy is helpful to the identification and categorization of sexism and misogyny in memes.

1.4 Thesis Structure

This thesis is organized as follows. Chapter 2 introduces related research on the detection of harmful content in both textual and multimodal content, with a focus on the strategies implemented to address the task in each format. Chapter 3 presents the datasets, preprocessing steps, and models implemented, including the multimodal baseline approach based on the visual and textual modalities, the conventional machine learning and the transformer-based models using the textual representation of memes, as well as the ensemble strategy. The chapter concludes with the evaluation metrics used for each task (binary and multi-label classification). Chapter 4 discusses the results of binary and multi-label classification tasks, in both in-domain and cross-dataset setups. It also provides a correlation analyses of the component models within the ensembles, as well as an analysis of the errors made by the ensemble strategies in binary classification. Finally, Chapter 5 concludes the project, highlights its limitations, and offers suggestions for future research.

Chapter 2

Related Work

The aim of this chapter is to provide an overview of the methods applied up to now in the detection of harmful content, both in relation to text and memes. Beginning with textual data, an overview of the approaches used are presented, focusing on the features and classifiers used in conventional machine learning, pre-trained models based on the transformer architecture, and ensemble strategies. Next, the approaches used for multimodal data, specifically in memes, are described. Both settings are described in relation to harmful memes in general and those that specifically target sexism and misogyny. The chapter concludes with a section that summarizes the goal of this thesis in relation to the research gap identified in the literature.

2.1 Detecting Harmful Content in Textual Data

Harmful content consists of a range of overlapping and intersecting phenomena, comprising various forms of expression that cause different harms. Among them, hate speech is the most widely recognized (Faris et al., 2016), which definition can very depending on the context in which it is applied and the purpose that it serves (Khurana et al., 2022). Hate speech can be defined as a communication that disparages a person or a group on the basis of characteristics such as race, ethnicity, gender, sexual orientation, nationality, or religion, among others (Nockleby (2000), as cited in Schmidt and Wiegand (2017)). While hate speech is a broad term that also encompasses sexism and misogyny, this thesis uses the term "harmful content" to refer to forms of hate speech in general.

Considering the textual modality, the detection of harmful content was initially approached using conventional machine learning relying on extensive feature engineering, which was later followed by the adoption of pre-trained transformer models. Both approaches are described below in relation to harmful content as well as sexism and misogyny.

2.1.1 Conventional Machine Learning Approaches

In conventional machine learning, models learn from data samples to make predictions on new observations. Harmful content detection is a supervised classification problem given that a model learns the correlation between the inputs and outputs using labelled training data samples to output a hard label (Chen et al., 2012; Zhou, 2021), i.e. either harmful or non-harmful. This section introduces the features and classifiers implemented for the detection of harmful content in general, as well as those specific to sexism and misogyny.

Features

As highlighted by Schmidt and Wiegand (2017), defining what constitutes an instance of harmful content can be challenging since it can be influenced by the characteristics of the domain in which it occurs, or its discourse or media context. What differentiates a positive from a negative instance cannot be always attributed to a single aspect and, therefore, a variety of features were implemented in previous research. The same authors categorized the most commonly implemented features for the detection of harmful content, and, based on their description (Schmidt and Wiegand, 2017), the most relevant to my research are summarized below.

Surface-level Features They refer to the presence of a word or character sequences, considering unigrams or larger n-grams, represented using TF (Term Frenquency) or TF-IDF (Term Frequency-Inverse Document Frequency) weighting schemes. As unusual spellings or rare words introduced by users might result in unknown tokens, character-level n-gram features can help address these spelling variations, and have even proven to be more effective than token-level n-grams (Mehdad and Tetreault, 2016). These features have been widely implemented for the identification of sexism and misogyny in tweets in both individual research (Anzovino et al., 2018; Frenda et al., 2019; Rodríguez-Sánchez et al., 2020), and as part of shared tasks (Fersini et al., 2018a,b). Due to their proven effectiveness, surface-level features are considered in this study for the identification and categorization of sexist and misogynous memes.

Lexical Resources This feature relates to the presence of specific words, slurs or insults, as well as verbs or adjectives that signal positive or negative instances of harmful content. However, these features alone are not sufficient when compared to surface-level or word generalization features. Regarding the detection of sexism and misogyny in tweets, one study (Frenda et al., 2019) implemented lexicons to capture aspects of aggressive messages against women, including feminity, vulgarity, sexuality, and parts of the human body, as well as a list of abbreviations and hashtags referring to stereotypes. For the EXIST 2021 shared task on sexism identification of tweets (Rodríguez-Sánchez et al., 2021), a few teams implemented Hurtlex (Bassignana et al., 2018), a multilingual lexicon of hate words, as a feature.

Linguistic Features These include part-of-speech (POS) tags or dependency relationships. For example, knowing that two words are syntactically related increases the likelihood that a statement expresses harmful content, compared to when the same words appear in a sentence without any syntactic connection. Markov et al. (2021) implemented POS tags to capture morpho-syntactic patterns, combined with function words to extract stylometric features, given their significance in reflecting the writing style of an author as a potential indicator of harmful content. In relation to misogyny identification and categorization of tweets, apart from POS tags, Anzovino et al. (2018) implemented the length of each instance since those labelled as sexual harassment were usually shorter, as well as the number of adjectives because stereotype and objectification tweets included more describing words, among others.

Sentiment Analysis Given that harmful content is typically assumed to carry a negative sentiment, sentiment analysis can be incorporated either as an auxiliary classification task in a multi-step approach or as a feature in a single-step classification. For example, the number of positive, negative or neutral words in a given instance can be used as a feature. In addition, polarity classifiers can be implemented, as harmful content often exhibits a high degree of negative polarity. Likewise, emotional features can be included as harmful content often exhibits emotional dimensions. Markov et al. (2021) incorporated emotion-conveying words found in an instance from the NRC lexicon (Mohammad and Turney, 2013) and their association to emotions (such as anger, fear, joy, among others) and sentiment (negative and positive). Moreover, emotion-conveying words were also combined with the above-described linguistic features containing POS and function words in the same study. To the best of my knowledge, emotion features have not been explored for the detection of sexism and misogyny in textual content.

In summary, surface-level features have proven effective in identifying sexist and misogynous textual content. However, to the best of my knowledge, stylometric and emotion-based features have not yet been explored for this purpose. Therefore, this research will implement surface-level features together with stylometric and emotion-based features to support the identification and categorization of sexism and misogyny in memes. Following Markov et al. (2021), stylometric and emotion-based features will be applied through a combination of lexical resources, linguistic and emotion features, which will be described in Chapter 3.

Feature Representation

To allow computers to process natural language, each data sample and its corresponding label must be converted into a numerical vector. Two common vectorization methods are TF and TF-IDF. In TF, each document is converted into a vector that counts the occurrences of words, with each entry indicating the number of times a particular word appears in the document. The length of the vector corresponds to the size of the vocabulary across all documents and those with similar words will produce similar vectors. TF-IDF, on the other hand, builds upon TF by adjusting the raw frequency counts. It assigns higher weights to words that appear in fewer documents, as they are more informative in distinguishing one document from another (Jurafsky and Martin, 2025). Both of these vectorization methods will explored in this thesis to represent the textual data for the classifier.

Classifiers

Among supervised machine learning algorithms for text classification, **Support Vector Machines (SVM)** have been widely used for the detection of harmful content (Schmidt and Wiegand, 2017; Zampieri et al., 2019; Markov and Daelemans, 2021). A similar trend is observed for the identification of sexism and misogyny in various studies (Frenda et al., 2019) and shared tasks, such as AMI (Automatic Misogyny Identification) at IberEval 2018 (Fersini et al., 2018b) and EXIST 2021 (Rodríguez-Sánchez et al., 2021). Furthermore, this classifier was reported to be the best performing approach when compared to other traditional machine learning methods, achieving an accuracy of 0.80 in misogyny identification and a macro F1 score of 0.38 for misogyny categorization (Anzovino et al., 2018). Other models explored for the detection of sexism and misogyny in tweets include Logistic Regression (LR) (Fersini et al., 2018a, 2020) and Random Forest (RF) (Anzovino et al., 2018; Rodríguez-Sánchez et al., 2021). Due to the wide application and success of SVM, this classifier will be used in this thesis for the identification and categorization of sexism and misogyny in memes.

2.1.2 Transformer Models

While conventional machine learning classifiers are useful for building lightweight systems, designing informative and effective features can be a complex task. However, deep learning models, particularly transformer-based architectures, support fine-tuning by enabling continued training on domain-specific data to adapt to a new domain or task (Jurafsky and Martin, 2025).

Given that **pre-trained models** can extract generalizations from large amounts of text, they have been widely applied to the identification of harmful textual content. Zampieri et al. (2019) highlighted that BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al., 2019) was the most commonly used model to address the identification of offensive language among the top 10 out of 104 teams at the SemEval-2019 Task 6. In the following edition of this shared task, SemEval-2020 Task 12 (Zampieri et al., 2020), which extended the challenge to also include other languages besides English, pre-trained models, such as BERT, RoBERTa (Liu et al., 2019) and other models based on the BERT architecture, were the most employed.

Focusing on sexism and misogyny, Rodríguez-Sánchez et al. (2020) reported that a fine-tuned multilingual version of BERT (mBERT) outperformed other conventional machine learning and deep learning models in the identification of sexist tweets in Spanish, achieving a macro F1 score of 0.64. BERT was also implemented by participants of AMI at EVALITA2020 shared task (Fersini et al., 2020). In the EXIST 2021 shared task (Rodríguez-Sánchez et al., 2021), most of the participants employed encoder-only architectures, such as BERT, RoBERTa, and their multilingual versions (mBERT, XLM-R (Conneau et al., 2020)) as well as BETO (Cañete et al., 2020), a Spanish version of BERT. This architecture outperformed the participants with conventional machine learning and deep learning systems in the sexism identification task, as it was used by the top 10 out of 33 teams. Furthermore, nearly 90% of the participants in the SemEval-2023 Task 10 on Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023) used a transformer architecture, with RoBERTa, DeBERTa (He et al., 2021, 2023), BERT, BERTweet (Nguyen et al., 2020), and DistilBERT (Sanh et al., 2020) being the most popular models.

Due to the popularity and effectiveness of BERT and RoBERTa across the different studies, this thesis will explore fine-tuning them to address the tasks in question. In addition, a version of BERTweet that originated at the EDOS shared-task (Al-Azzawi et al., 2023) will be explored. These models will be described in Chapter 3.

2.1.3 Ensemble Strategies

Previous research that adopted ensemble strategies to address the detection of harmful textual content showed that they outperform individual component models, in particular when combining conventional machine learning and deep learning approaches (van Aken et al., 2018; Markov et al., 2021, 2022). van Aken et al. (2018) implemented deep learning strategies such as birectional RNNs and CNNs combined with LR in an ensemble with gradient boosting decision trees. Their ensemble strategy determined the most effective classifier for each comment by observing its features and learning to weight and select the optimal classifier based on specific feature combinations, achieving a macro F1 of 0.79 in the classification of toxic comments. Markov et al. (2021) combined CNN and BERT with an SVM model with features related to the writing style and the emotion information in harmful content, using a hard majority voting ensemble. This system achieved a macro F1 score of 0.60 and 0.73 in the identification of harmful content in English and Dutch, respectively, when tested in a cross-domain setting. Markov and Daelemans (2021) implemented an ensemble of this SVM approach with BERT and RoBERTa through majority voting and confirmed that the combined predictions outperformed those of the individual component models in both in-domain and cross-domain settings. They also found that the SVM approach produced highly uncorrelated predictions compared to those made by the encoder-only models. The same SVM approach was implemented in the study by Markov et al. (2022), which combined BERTje and RobBERT in a grading boosting ensemble to address the detection of harmful content in Dutch, both in in-domain and cross-domain setups, resulting in a macro F1 score of 0.78 and 0.77 in the in-domain setting, and 0.59 and 0.74 in the cross-domain setting.

Ensembles have also been widely used for the identification of sexism and misogyny in text, where they have also proven to be beneficial. The best result for the identification of misogyny and aggressive behaviour in Italian at the AMI shared task held at EVALITA 2020 (Fersini et al., 2020) was achieved by an ensemble of custom BERT models that were fine-tuned with additional data, which scored a macro F1 of 0.74 (Lees et al., 2020). Participants in the EXIST 2021 shared task (Rodríguez-Sánchez et al., 2021) also employed ensemble strategies that combined different systems. The best performing team in sexism identification and categorization combined different BERT-based models, resulting in a macro F1 score of 0.78 for sexism classification and 0.58 for sexism categorization. Regarding conventional machine learning models, one team combined RF, LR and SVM in an ensemble for sexism categorization while another team implemented an ensemble strategy of both conventional machine learning and transformer-based models, namely XGBoost (eXtreme Gradient Boosting), SVM, RoBERTa for English and mBERT for Spanish, achieving a macro F1 score of 0.72.

Considering that ensembles have proven beneficial for the detection of textual harmful content, including sexism and misogyny, this thesis aims to investigate the use of an ensemble of conventional and deep learning models for the detection of sexist and misogynous memes. A key contribution is the incorporation of cross-dataset evaluation, which, to the best of my knowledge, has not been yet explored in this context. The approach takes inspiration from ensembles that combine conventional machine learning and transformer-based models and have outperformed the individual models in harmful content detection. Moreover, since an SVM classifier with stylometric and emotion-based features has shown promising results in this domain when incorporated in an ensemble by producing uncorrelated predictions, it could also be valuable for the detection and categorization of sexism and misogyny. These specific features have not yet been explored for these tasks, to the best of my knowledge. Therefore, the motivation of this thesis builds on the research of van Aken et al. (2018); Markov et al. (2021); Markov and Daelemans (2021) and Markov et al. (2022), aiming to implement an ensemble of both shallow and deep learning models to address the detection of sexism and misogyny but applied to memes.

2.2 Detecting Harmful Content in Multimodal Data

Compared to the textual approaches, the work to address the identification of harmful content in multimodal data has been emerging. Different shared tasks have been implemented to address this task particularly focusing on memes, such as Hateful Memes Challenge (Kiela et al., 2020) and Multimodal Hate Speech Event Detection (Thapa et al., 2024). In addition, some approaches and shared tasks have specifically targeted the identification and categorization of sexism and misogyny in memes, namely Sexist MEME Detection (Fersini et al., 2019), EXIST 2024 (Plaza et al., 2024a), and MAMI (Multimedia Automatic Misogyny Identification) (Fersini et al., 2022). The systems developed in these works vary in their architecture, ranging from conventional machine learning approaches and pre-trained models, to more complex architectures and ensemble strategies, which are explained in the below sections.

2.2.1 Approaches for Detecting Harmful Memes

In the Hateful Memes Challenge (Kiela et al., 2020), the task was to classify memes as hateful or not, using the area under the receiver operating characteristic curve (ROC AUC) and accuracy as evaluation metrics. The authors introduced different baselines with unimodal and multimodal pre-trained models. The latter were pre-trained either unimodally or multimodally. The unimodal models included image encoders (Image-Grid based on ResNet-152, and Image-Region based on Faster-RCNN) and BERT for text data, with BERT achieving the highest performance with an ROC AUC of 0.69 and accuracy of 0.63. The multimodal models pre-trained on unimodal data used (1) simple fusion methods and (2) advanced multimodal models. The simple fusion methods included Late Fusion, a model that combined ResNet-152 for images and BERT for text by averaging their outputs, and Concat BERT, which concatenated the features from the models implemented in Late Fusion and trained a multi-layer perceptron (MLP) on top to make predictions. The advanced multimodal models were MMBT-Grid and MMBT-Region, supervised multimodal bitransformers that used features from Image-Grid or Image-Region; and versions of VilBERT and Visual BERT pre-trained on unimodal data. Among these models, MMBT-Region performed best with an ROC AUC of 0.74 and accuracy of 0.68. Finally, the multimodally pre-trained Visual BERT COCO outperformed ViLBERT CC with an ROC AUC of 0.75 and accuracy of 0.69. The preliminary results indicated that models performed better with more advanced fusion methods such as early fusion, outperforming middle and late fusion approaches.

Wang and Markov (2024a,c,b) implemented a **multimodal model** that resulted in the winning approach of multiple shared tasks in English as well as other languages: Multimodal Hate Speech Even Detection (Thapa et al., 2024), which targeted multimodal hate speech related to the Russo-Ukrainian War and its targets, DIMEMEX (Detection of Inappropriate Memes from Mexico) (Jarquín-Vásquez et al., 2024), which aimed to detect fine-grained types of hate speech in Mexican Spanish memes, and ArAIEval (Hasanain et al., 2024) which focused on multimodal propagandistic meme classification in Arabic. The multimodal approach concatenated the embeddings from a Swin Transformer-based visual model and a pre-trained language models using the MLP fusion model (Shi et al., 2021), achieving macro F1 scores of 0.87 for hate speech detection and 0.80 for hate speech target detection in the Multimodal Hate Speech Event Detection Challenge 2024 (Wang and Markov, 2024a). At the DIMEMEX shared task, the system resulted in an F1 score of 0.58 for classifying memes into hate speech, inappropriate, or harmless categories, and an F1 score of 0.44 for the fine-grained classification of hateful memes (Wang and Markov, 2024c). This architecture achieved a macro F1 score of 0.80 at ArAIEval, ranking third on the leaderboard for the multimodal subtask (Wang and Markov, 2024b).

Regarding the research related to the detection of sexism and misogyny in memes, to the best of my knowledge, Fersini et al. (2019) introduced the first work to address this problem. The task was tackled using both unimodal and multimodal approaches, with models such as SVM, Naïve Bayes, decision trees and k-nearest neighbours (kNN). The unimodal approaches were trained with either textual or visual features. The textual features were based on a TF weighting scheme applied to the text in the memes, while the visual features included colour, photographic and aesthetic features, features related to visual perception, and features related to semantic concepts such as the percentage of visible skin and the number of faces. While SVM performed best on textual data, achieving a macro F1 score of 0.76, kNN was the best visual approach, with a macro F1 score of 0.62. However, kNN struggled to detect sexist memes, indicating that handcrafted visual features were poor indicators of sexism in meme images. On the other hand, the multimodal approaches combined textual and visual representations using early and late fusion strategies. Late fusion outperformed early fusion, with a macro F1 of 0.76 using an SVM, compared to 0.69 with a decision tree. The authors found that a joint representation of textual and visual features was not sufficient to capture the complexity of sexist memes.

The MAMI Shared Task (Fersini et al., 2022), introduced at SemEval 2022, focused on the detection of misogynous memes in English. It consisted of two subtasks: misogyny detection and misogyny categorization. Most of the teams implemented pre-trained models, with BERT-based models being the most commonly used for the textual modality, and models based on VisualBERT the most used for the visual modality. The evaluation metrics were macro F1 for the binary classification task and weighted F1 for the multi-label classification task. For the misogyny categorization task, the system proposed by Hakimov et al. (2022) achieved a weighted F1 score of 0.73. It consisted of a neural model that used CLIP to extract textual and visual features, which were processed separately with LSTM for textual, and a fully connected layer for visual features. Their outputs were fed into separate dropout layers, concatenated, and passed through another fully connected layer. The final vector representation was passed to separate sigmoid functions to predict class probabilities for each subtask.

To the best of my knowledge, only the detection of misogyny was further researched after the shared task, with Rehman et al. (2025) achieving a macro F1 of 0.88 on the test set by implementing a multimodal context-aware attention-based model. Their approach consisted of three modules: a Multimodal Attention Module (MANM), a Graph-based Feature Reconstruction Module (GFRM), and a Content-specific Feature Learning Module (CFLM). The MANM captured the interaction between regions of an image and words in the accompanying text by integrating the global context of the image. It applied cross-attention to produce cross-modal representations for both modalities, followed by multi-head self-attention to focus on relevant information within these features. The GFRM reconstructed unimodal features to enhance the representation of each modality, while the CFLM used additional content-specific features, including a misogyny-specific lexicon and toxicity indicators for text, as well as image caption embeddings and NSFW features to identify indecent content in images. The classification module fused features from the three modules to create a unified representation, which was used to generate the final predictions. During testing, the model incorporated additive perturbation-based test-time augmentations to improve generalization to variations in the test data that might not have been present in the training data.

As for sexism, the 2024 edition of EXIST (Plaza et al., 2024a,b) focused on the identification of sexism, source intention identification, and sexism categorization in tweets and memes in English and Spanish. The shared task adopted the learning with disagreement (LwD) paradigm, in which systems were developed to learn from different perspectives, biases, or interpretations from the annotators. Teams could participate using systems that implemented the LwD setup, outputting a probability for the label(s) in each subtask, as well as a traditional setup, in which systems produced a hard label. The evaluation metrics varied depending on the system output. Focusing on systems that produced a hard label, the official metrics were ICM, as well as a normalized version of ICM and F1. In sexism identification, the F1 score for the positive class was reported, while macro F1 was used for the remaining subtasks. While most teams approached the task as multimodal, the best performances came from text-based models, with encoding-based transformers being the most commonly used for this modality. Multimodal approaches followed, with most systems using CLIP to process images. However, the approaches that only included CLIP led to poor results (Plaza et al., 2024b).

Regarding what is relevant for this thesis, I will concentrate on the systems that produced a hard label for the English subset of memes and the subtasks of sexism identification and categorization. The best approach for sexism identification was implemented by Ma and Li (2024), prompting GPT-4 to generate descriptions of the memes together with a label indicating whether the meme was sexist or not. These output labels from the GPT-4 model in a zero-shot setting achieved the best results with an ICM score of 0.34 and an F1 score of 0.78 for the sexist label. The best-performing approach for sexism categorization was implemented by Menárguez Box and Torres Bertomeu (2024), which resulted in an ICM score of -0.70 and a macro F1 of 0.43. They augmented the training data for some categories by generating three new samples for each instance using BERT contextual embeddings. The model used was a BERTweet-large-sexism-detector model (Al-Azzawi et al., 2023), fine-tuned with the dataset from SemEval-2023 Task 10 on Explainable Detection of Online Sexism (Kirk et al., 2023). The task was addressed with a model trained on both English and Spanish for the sexism label, and separate models to predict each category in English.

While different methods have been used for the detection of sexist and misogynous memes, ranging from conventional machine learning to pre-trained transformer models and vision-language models (VLM), this thesis will implement a multimodal model that processes both image and text. This will be done by taking inspiration from the approach introduced by Wang and Markov (2024a,b,c) due to its high performance on different shared tasks. This multimodal model, with its architecture described in Chapter 3, will also serve as a baseline model to compare the performance of the ensemble strategy.

2.2.2 Ensemble Strategies for Multimodal Harmful Content Detection

Ensemble strategies have also been implemented for the detection of harmful memes. Considering the Hateful Memes Challenge, the first-place winning solution (Zhu, 2020) employed an ensemble of four VLM: VL-BERT, UNITER, VILLA-ITM, and ERNIE-

Vil. Apart from incorporating both the image and text as input, the models also included other features such as entity, race and gender. This approach achieved a ROC AUC of 0.84 and an accuracy of 0.73.

The winning approach for misogyny identification at MAMI Shared Task (Zhang and Wang, 2022) implemented an ensemble of pre-trained models, boosting, and rule-based adjustment, achieving a macro F1 score of 0.83. The approach involved three steps: first, an XGBoost classifier was trained using image features from CLIP, a UNITER model was fine-tuned on the paired image and meme text, and additional external datasets were used to fine-tune a BERT model. Then, the confidence zone of the XGBoost predictions were adjusted with those from UNITER and BERT. Finally, the predictions for the binary classification and multi-label classification tasks were mutually adjusted to take advantage of their logical inference relationship.

In the misogyny categorization task, two other teams achieved the highest weighted F1 score of 0.73 (as did the system described in Section 2.2.1). The winning approach for misogyny categorization (Zhang and Wang, 2022) mentioned above was one of them together with the system proposed by Zhi et al. (2022). They implemented an ensemble strategy that combined two multimodal models: a fine-tuned image classification model based on ConvNext to extract visual features, and either RoBERTa or DeBERTa to extract textual features. The textual and visual features were concatenated and fed into a two-layer fully connected neural network for classification. The outputs of the two multimodal models were fused using weighted voting, with a threshold for selecting the positive classification label.

While ensemble strategies have also proven to be effective for the classification of harmful memes, the combination of deep learning and conventional machine learning models was not explored. Considering that SVM showed to be useful for the classification of textual harmful content (where it provided uncorrelated predictions) as well as memes, and that features addressing the writing style and emotion were not implemented for the identification and categorization of sexism and misogyny, in particular in memes, this thesis aims to address this gap by applying the SVM approach introduced by Markov et al. (2021) in an ensemble. Apart from SVM, the models to be combined in an ensemble to address the identification and categorization of sexist and misogynous memes are: pre-trained transformer-based models, which have shown strong performance for the identification of sexism and misogyny both in textual and multimodal content, and the multimodal model introduced by Wang and Markov (2024a,c,b), which achieved state-of-the-art performance for the classification of harmful memes. This multimodal approach, as a result, will serve as a baseline to determine whether the proposed ensemble could outperform it. Furthermore, the multimodal architecture has not been used in an ensemble for the detection of harmful multimodal content yet.

2.3 Research Focus

In summary, pre-trained transformer-based language and vision models have been widely used to tackle the identification of harmful content in memes, as well as the identification and categorization of misogyny in MAMI, and sexism in EXIST 2024. Pre-trained language models also showed to be the best systems to classify textual content, either harmful, sexist or misogynous. Across these efforts, the best performance was achieved when using ensemble strategies that combined different models, as they outperformed the component models. Moreover, a multimodal model combining text and vision embeddings resulted in state-of-the-art performance in the detection of harmful memes.

While ensemble strategies that incorporate deep learning together with shallow approaches have shown promising results for identifying harmful textual content, they have not been implemented to address the identification and categorization of sexism and misogyny in memes. To address this gap, an ensemble of conventional machine learning and transformer-based models will be implemented to address these tasks. Since stylometric and emotion-based features have not been explored to tackle the classification of sexism and misogyny yet, an SVM that incorporates these hand-crafted and informative features will be combined in an ensemble. This SVM has demonstrated promising results for detecting harmful content in text when incorporated in an ensemble, given that it produced uncorrelated predictions compared to those made by encoder-only models. In addition, as the multimodal model proposed by Wang and Markov (2024a,b,c), which combines transformer-based vision and language models, has not been explored for the identification and categorization of sexism and misogyny in memes, it could serve as a baseline experiment to explore whether the proposed ensemble would be more beneficial for these tasks. Thus, the ensemble strategy will incorporate an SVM with stylometric and emotion-based features, a pre-trained transformer-based language model, for which BERT, RoBERTa, and a BERTweet fine-tuned for sexism (Al-Azzawi et al., 2023) will be explored, and a multimodal model.

A key contribution of this thesis is the implementation of a cross-dataset evaluation setup to assess the ability of the models to generalize beyond their training data, reflecting the challenges faced in real-world applications. Therefore, this thesis aims to answer whether an ensemble strategy that incorporates both conventional and deep learning models are beneficial for the identification and categorization of sexism and misogyny in memes. This will be evaluated across in-domain and cross-dataset settings.

Chapter 3

Methodology

This chapter describes the methodology applied in this thesis. It starts with the meme datasets used to address the binary and multi-label classification tasks: MAMI (Fersini et al., 2022) and EXIST 2024 (Plaza et al., 2024a,b). The implemented preprocessing steps are described, followed by the experimental setup, which describes the approaches to the multi-label classification task. Next, the component models and the ensemble strategies are explained, along with the evaluation metrics.

3.1 Datasets

3.1.1 MAMI

This dataset was a contribution of SemEval-2022 Task 5, Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022), where a meme was to be classified as misgoynous or not, and the type of misogyny was to be identified. The memes were collected from social media platforms such as X (formerly Twitter) and Reddit, and websites dedicated to the creation and sharing of memes, such as 9GaG, Knowyourmeme and Imgur. Beside the memes, the meme text was also part of the dataset, which consisted of a total of 11,000 memes split into 10,000 for training and 1,000 for testing. Inter-annotator agreement (IAA) was reported with Fleiss-k coefficient (Fleiss, 1971). IAA for misogyny identification resulted in a moderate agreement of 0.58 with traditional Fleiss-k, suggesting that it was a simple task for humans. For misogyny categorization, however, the fair agreement of 0.34, which was computed with the MASI (Jaccard) index (Passonneau, 2006), denoted a difficult task instead (Fersini et al., 2022).

Misogyny Identification

In the binary classification task, memes were classified as either misogynous or non-misogynous. A description of each category is provided below, based on the MAMI Shared Task overview publication (Fersini et al., 2022).

- **Misogynous**: The meme conceptually depicts an offensive, sexist or hateful scene (whether weak or strong, implicitly or explicitly) targeting a woman or a group of women. Misogyny can be expressed in the form of shaming, stereotype, objectification and/or violence.

Binary Label	Total	Training		Test	
		#	%	#	%
Misogynous	**5,500**	5,000	50	500	50
Non-misogynous	**5,500**	5,000	50	500	50
Total	**11,000**	**10,000**	**100**	**1,000**	**100**

Table 3.1: Statistics of the MAMI training and test sets for misogyny identification.

- **Non-misogynous**: The meme does not depict any form of hatred of women.

The distribution of misogynous and non-misogynous memes was balanced, with 5,000 memes per category in the training set and 500 memes in the test set, as shown in Table 3.1 and Figure 3.1 (top).

Misogyny Categorization

In the multi-label classification task, a meme was classified into one or more overlapping fine-grained classes. The types of misogyny considered in MAMI were shaming, stereotype, objectification, and violence. A description of each category is provided below.

- **Shaming**: Memes that insult and offend women because of characteristics of their body or personality. These memes are related to denigrating the physical appearance of women (body shaming).

- **Stereotype**: Memes that aim to represent a fixed idea or set of characteristics assigned to women. These memes convey the image of women according to their role in society (i.e., role stereotyping), their personality traits and domestic behaviours (i.e., gender stereotyping), or fixed ideological characteristics related to women's rights (i.e., feminism stereotype).

- **Objectification**: Memes that see or treat women as objects. These memes typically express an exaggerated appreciation of the physical appearance of women, depicting them either as sexual objects or as human beings without value as individuals.

- **Violence**: Memes that indicate physical or verbal violence. These memes aim to depict violence against women or allude to the intent to physically assert power over them.

Table 3.2 presents the distribution of the fine-grained categories. Since memes could belong to more than one fine-grained misogynous category, their distribution was unbalanced, with a total of 7,239 memes across all categories in the training set, and 997 in the test set. An exploration of the distribution of fine-grained labels revealed that a few memes labelled as non-misogynous contained fine-grained category labels, which occurred for the shaming and objectification classes, with 3 and 1 non-misogynous memes, respectively.

Focusing on the fine-grained classes across misogynyous memes, they exhibited significant imbalance and were underrepresented when considered individually. The stereotype category was the most frequent fine-grained class, with 2,810 memes labelled

Binary Label	**Fine-grained Label**	**Total**	**Training**		**Test**	
			#	%	#	%
Misogynous	Shaming	1,417	1,271	25.4	146	29.2
	Stereotype	3,160	2,810	56.2	350	70.0
	Objectification	2,549	2,201	44.0	348	69.6
	Violence	1,106	953	19.1	153	30.6
Non-misogynous	Shaming	3	3	0.06	n/a	
	Stereotype	0	0	-		
	Objectification	1	1	0.02		
	Violence	0	0	-		
Total		**8,236**	**7,239**	-	**997**	-

Table 3.2: Statistics of the MAMI training and test sets for misogyny categorization. Percentages considered in terms of the binary label.

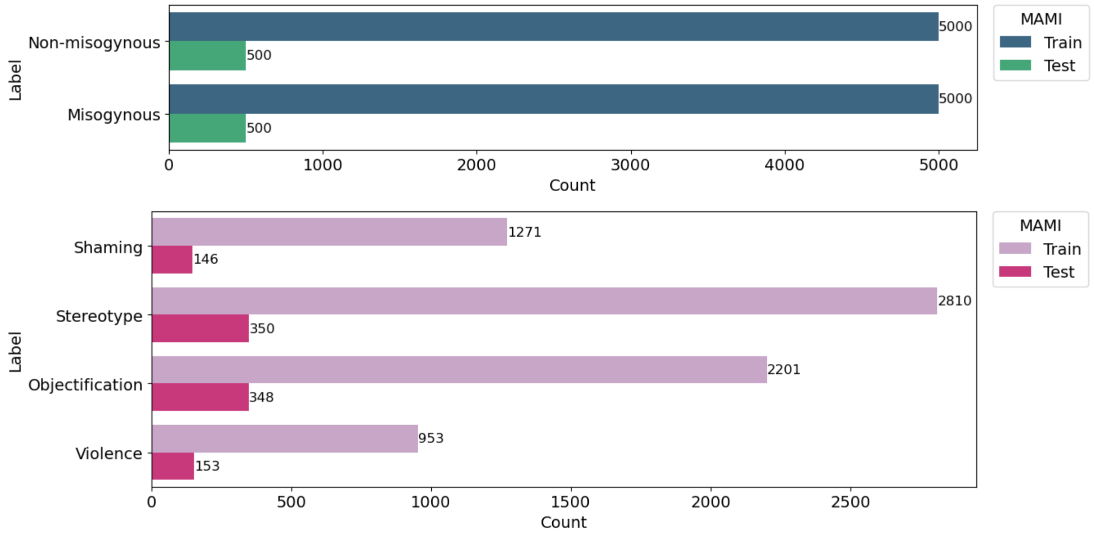


Figure 3.1: Distribution of binary (top) and fine-grained classes when meme is misogynous (bottom) in the MAMI training and test sets.

as such (28.1% of the training dataset and 56.2% of the misogynous instances). This was followed by objectification, with a total of 2,202 memes (22.0% of the dataset and 44.0% of the misogynous memes), and shaming, with 1,274 memes (12.7% and 25.4%, respectively), both of which included labels across misogynous and non-misogynous instances. The least represented category was violence, which appeared in 953 memes (9.5% of the dataset, and 19.1% of the misogynous memes). A similar pattern was observed in the test set. Stereotype remained the most common category, with 350 out of 500 misogynous memes (35.0% of the total test set and 70.0% of the misogynous instances), while the least frequent class was shaming, with 146 memes (14.6% and 29.2%). Violence followed with 153 memes (15.3% and 30.6%) and objectification with 348 memes (34.8% and 69.6%). Figure 3.1 (bottom) presents the distribution of the fine-grained classes of misogynous memes in both training and test sets.

Binary Label	Training			
	#	%	# with Cat	%
Sexist	965	56	958	48
Non-sexist	743	43	743	37
Total	1,708	100	**1,701**	**100**

Table 3.3: Statistics of the EXIST 2024 training set for sexism identification. # with Cat: Number of memes with fine-grained classes.

3.1.2 EXIST 2024

The dataset originated in the EXIST Shared Task at CLEF 2024 (Plaza et al., 2024a,b). It contained 5,097 memes in English and Spanish, collected from Google Images and annotated by six annotators. As the MAMI dataset described above, this dataset also included the text contained in the memes. While the dataset was balanced between the two languages, this thesis used the English portion, which contained 2,010 in the training dataset, and 513 in the test set; however, the test set labels were not publicly available. In order to create hard labels for the dataset, the authors applied a probabilistic threshold to the labelled training set. For sexism identification, the class annotated by more than three annotators was selected as the hard gold label, while for sexism categorization, the class annotated by more than one annotator was selected. As a result, the size of the training dataset was reduced whenever these thresholds were not met, which led to more than 300 English memes missing a hard label.

Sexism Identification

The categories addressed in binary classification are described below based on the EXIST 2024 Shared Task overview (Plaza et al., 2024b) and annotation guidelines.

- **Sexist**: The Oxford English Dictionary defines sexism as “prejudice, stereotyping or discrimination, typically against women, on the basis of sex”. Sexism encompasses any form of oppression or prejudice against women due to their sex. This discrimination can stem from different beliefs, such as stereotypes, the belief that men are superior to women, or an irrational hatred of women, commonly referred to as misogyny. The latter represents a more extreme, hate-driven form of sexism. The meme can be sexist itself, describe a sexist situation or criticize a sexist behaviour.

- **Non-sexist**: The meme does not prejudice, underestimate, or discriminate against women.

While the original English meme training dataset contained a total of 1,708 memes labelled for sexism, seven sexist memes did not contain hard fine-grained labels for sexism categorization. These memes were disregarded, resulting in 1701 memes containing both binary and fine-grained labels. The distribution of the classes for sexism identification is presented in Table 3.3 and Figure 3.2 (top). This dataset presented more sexist memes than non-sexist ones, with 56% vs 43%, respectively.

Fine-grained Label	**Training**	
	#	%
Ideological and inequality	408	42.3
Misogyny and non-sexual violence	180	18.7
Objectification	459	47.6
Sexual violence	213	22.1
Stereotyping and dominance	480	49.7
Total	**1,701**	**-**

Table 3.4: Statistics of the EXIST 2024 training set for sexism categorization. Percentages considered in terms of the *sexist* label.

Sexism Categorization

In EXIST 2024, the types of sexism that memes were classified into in multi-label classification included ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence, which are described below.

- **Ideological and inequality**: Memes that discredit the feminist movement to devalue, belittle and defame the struggle of women in all areas. Also included are memes that reject inequality between men and women, or present men as victims of gender-based oppression.
- **Stereotyping and dominance**: Memes that express false ideas about women that suggest they are more suitable or inappropriate for certain tasks. Also included are any memes that imply that men are somehow superior to women.
- **Objectification**: Memes where women are presented as objects apart from their dignity and personal aspects. Also included are memes that assume or describe certain physical qualities that women must have to fulfil traditional gender roles.
- **Sexual violence**: Memes where sexual suggestions, requests or harassment of a sexual nature are made.
- **Misogyny and non-sexual violence**: Memes where expressions of hatred and violence towards women are contained.

The distribution of the fine-grained categories of sexist memes in the training dataset is provided in Table 3.4. In this dataset, only memes labelled as sexist were annotated for the type of sexism. The most frequent categories were *stereotyping and dominance* and *objectification*, with 480 memes (49.7% of sexist memes) and 459 memes (47.6%), respectively. They were followed by *ideological and inequality* with 408 memes (42.3%) and *sexual violence* with 213 memes (22.1%). The least frequent category was *misogyny and non-sexual violence*, with 180 memes (18.7% of the dataset). Figure 3.2 (bottom) presents the distribution of the fine-grained classes in EXIST 2024 dataset.

3.1.3 Stratified Sampling

Due to the multi-label nature of the datasets, stratified sampling was implemented to split the datasets into training, development, and test sets. Since the original EXIST 2024 test set was not publicly available, the training set was used to generate the splits

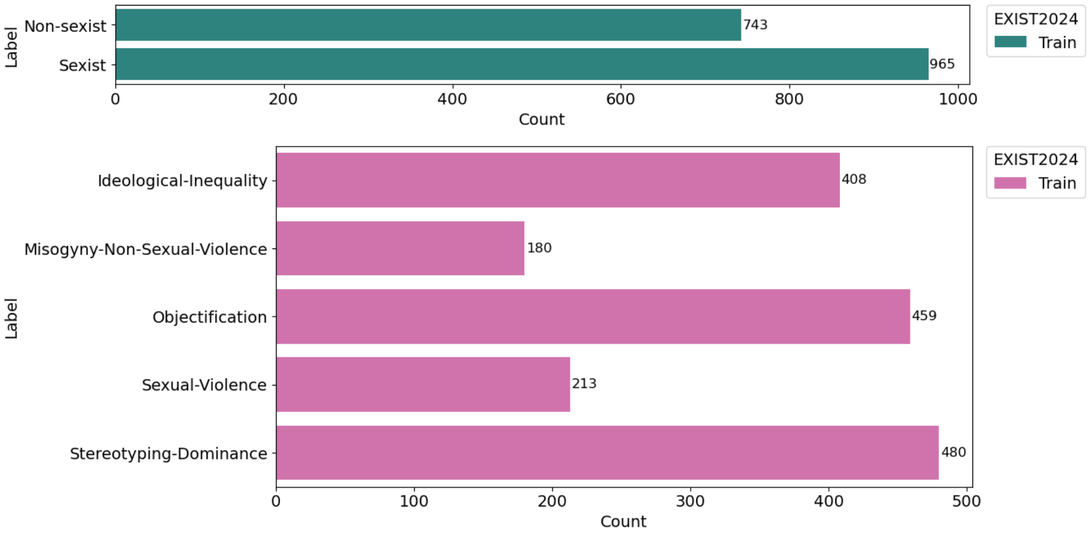


Figure 3.2: Distribution of binary (top) and fine-grained classes when meme is sexist (bottom) in the EXIST 2024 dataset.

for all experiments. Stratified sampling is a method that accounts for disjoint groups within a population and ensures that each sample reflects the overall class distribution by preserving the proportion of instances from each subgroup as in the complete dataset (Sechidis et al., 2011). The method applied was *MultilabelStratifiedShuffleSplit* from the iterative-stratification[1] library with the random seed set to 0 for reproducibility. This method was implemented differently to prepare the datasets for in-domain and cross-dataset experiments, as described below.

In-Domain Experiments

For the in-domain experiments, which were carried out on the complete, individual datasets, each dataset was split using stratified sampling. Given that the MAMI dataset already contained a test split, it remained unchanged and the training dataset was divided into training and development subsets with a 90%-10% split. The EXIST 2024 training dataset was split into 80% training, 10% development, and 10% test sets.

The final class statistics of the MAMI dataset splits are shown in Table 3.5 for the binary classes and in Table 3.6 for the fine-grained classes. When examining the percentage distribution, the fine-grained classes resulted in similar proportions across the training and development sets, as was expected with the applied stratifying method. Figure 3.3 illustrates the class distribution for both binary and fine-grained classes in the MAMI dataset splits after stratified sampling.

The statistics of the EXIST 2024 meme dataset after stratified sampling are presented in Tables 3.7 and 3.8 for the binary and fine-grained classes, respectively. Since all splits were derived from a single dataset, the class distribution remained consistent across the splits, in particular for the fine-grained categories. The distribution of both binary and fine-grained classes is shown in Figure 3.4.

[1] https://github.com/trent-b/iterative-stratification

Binary Label	**Train**		**Dev**		**Test***	
	#	%	#	%	#	%
Misogynous	4,483	49.8	517	51.7	500	50
Non-misogynous	4,517	50.2	483	48.3	500	50
Total	9,000	100	1,000	100	1,000	100
Original total	10,000				1,000	

Table 3.5: MAMI binary class distribution after stratified sampling for in-domain experiments. *Test from original dataset.

Binary Label	**Fine-grained Label**	**Train**		**Dev**		**Test***	
		#	%	#	%	#	%
Misogynous	Shaming	1,132	25.3	139	26.9	146	29.2
	Stereotype	2,519	56.2	291	56.3	350	70.0
	Objectification	1,978	44.1	223	43.1	348	69.6
	Violence	846	18.9	107	20.7	153	30.6
Non-misogynous	Shaming	3	0.06	0	-	n/a	
	Stereotype	0	-	0	-		
	Objectification	0	-	1	0.2		
	Violence	0	-	0	-		
Total		6,478	-	762	-	997	-
Original total		7,239				997	

Table 3.6: MAMI fine-grained class distribution after stratified sampling for in-domain experiments. *Test from original dataset.

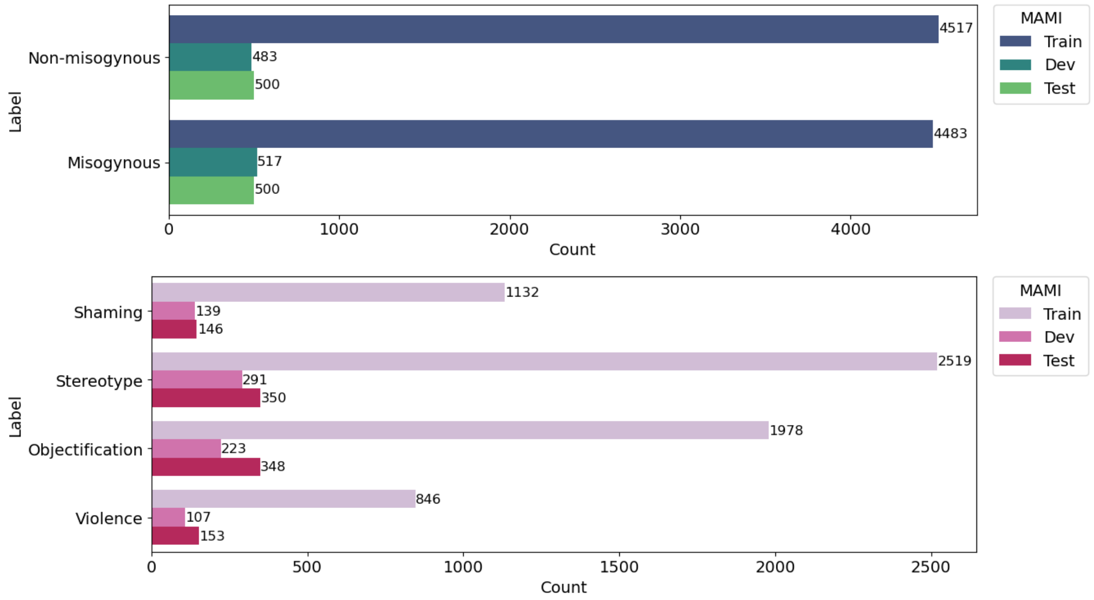


Figure 3.3: MAMI dataset splits after stratified sampling for in-domain experiments. Distribution of binary (top) and fine-grained classes when meme is misogynous (bottom).

Binary Label	**Train**		**Dev**		**Test**	
	#	%	#	%	#	%
Sexist	783	57	80	52	95	56
Non-sexist	594	43	73	48	76	44
Total	1,377	100	153	100	171	100
Original total	1,701					

Table 3.7: EXIST 2024 binary class distribution after stratified sampling for in-domain experiments.

Fine-grained Label	**Train**		**Dev**		**Test**	
	#	%	#	%	#	%
Ideological and inequality	333	42.5	36	45.0	39	41.1
Misogyny and non-sexual violence	147	18.8	17	21.2	16	16.8
Objectification	375	47.9	41	51.2	43	45.3
Sexual violence	186	23.8	11	13.8	16	16.8
Stereotyping and dominance	401	51.2	39	48.8	40	42.1
Total	1,442	-	144	-	154	-
Original total	1,740					

Table 3.8: EXIST 2024 fine-grained class distribution after stratified sampling for in-domain experiments.

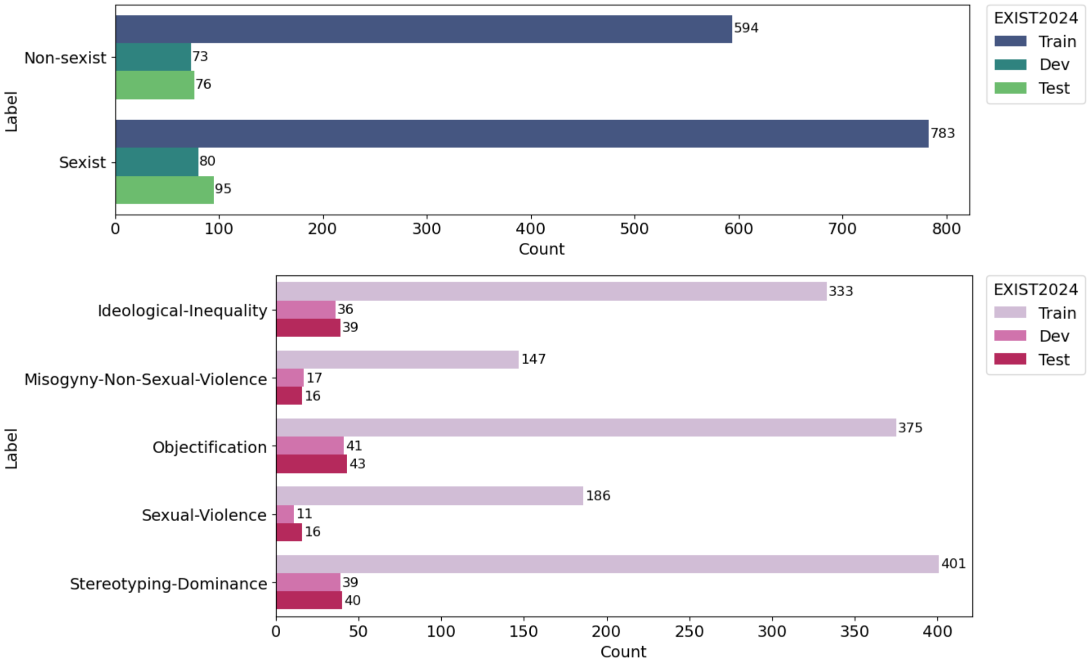


Figure 3.4: EXIST 2024 dataset splits after stratified sampling for in-domain experiments. Distribution of binary (top) and fine-grained classes when meme is sexist (bottom).

Binary Label	$\mathbf{T_o}$	**Train**		**Dev**		**Test ($\mathbf{T_o}$)**	
		#	%	#	%	#	%
Sexist	725	658	53	67	48	76	50
Non-sexist	667	594	47	73	52	76	50
Total	1,392	1,252	100	140	100	152	100
Total $\mathbf{T_o}$	1,544						

Table 3.9: EXIST 2024 binary class distribution after stratified sampling for cross-dataset experiments. T_o: Total with overlapping classes.

Fine-grained Label	$\mathbf{T_o}$	**Train**		**Dev**		**Test ($\mathbf{T_o}$)**	
		#	%	#	%	#	%
Objectification	416	377	57.3	39	58.2	43	56.6
Sexual violence	197	175	26.6	22	32.8	16	21.1
Stereotyping and dominance	440	400	60.8	40	59.7	40	52.6
Total	1,053	952	-	101	-	99	-
Total $\mathbf{T_o}$	1,152						

Table 3.10: EXIST 2024 fine-grained class distribution after stratified sampling for cross-dataset experiments. T_o: Total with overlapping classes.

Cross-Dataset Experiments

To conduct cross-dataset experiments, it was necessary to keep only the overlapping fine-grained categories for multi-label classification across datasets. Regarding the binary labels, the negative label remained unchanged in all splits as those memes did not contain any fine-grained classes. For the positive label, the memes with overlapping fine-grained labels were kept.

The fine-grained classes deemed similar in both datasets were objectification and stereotype from MAMI, which correspond to objectification and stereotyping and dominance in EXIST 2024. The violence class from MAMI was selected as a third category. Although it was challenging to directly match this category with one of the violence-related classes in EXIST 2024 dataset based on the definitions provided by the authors, sexual violence was considered the closest equivalent after a careful examination of several memes.

The training and validation sets obtained after stratified sampling for the in-domain setup described above were combined to remove the non-overlapping fine-grained classes. Due to the multi-label nature of the data, this filtering process resulted in some memes lacking any fine-grained labels. Thus, instances that were positive at the binary level but did not include any of the overlapping fine-grained categories were removed. The resulting dataset was then re-split with stratified sampling into new training and development sets with a 90%-10% ratio. The same procedure was applied to the test sets, where non-overlapping classes and instances without any fine-grained labels were were dropped.

The EXIST dataset was processed first due to its smaller size, resulting in a distribution of 667 non-sexist and 725 sexist memes. Of sexist memes, 416 belonged to the objectification class, 197 to the sexual violence class, and 440 to the stereotyping and dominance class in the joint training dataset before it was split into the new training and development sets as described above. The statistics for each split are provided in

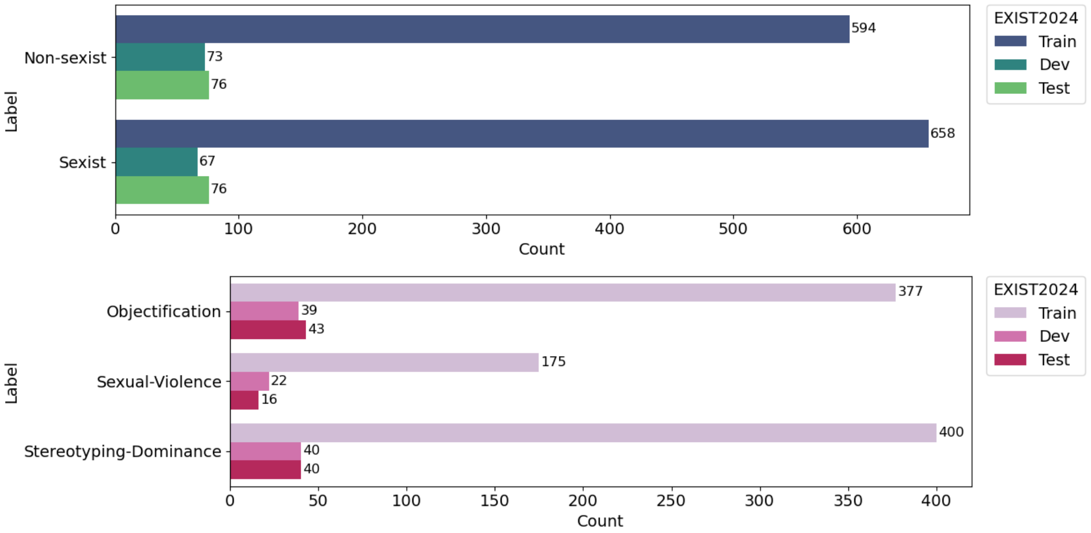


Figure 3.5: EXIST 2024 dataset splits after stratified sampling for cross-dataset experiments. Distribution of binary (top) and fine-grained classes when meme is sexist (bottom).

Table 3.9 and 3.10, for binary and fine-grained labels respectively, including the distribution before spitting. This resulted in a nearly balanced binary distribution across all splits, with 53% sexist and 47% non-sexist memes in the training set, 48%-52% in the development set and 50%-50% in the test set. The proportions of fine-grained classes remained consistent across the splits. The sexual violence category, being the least represented among the overlapping classes, led to a slightly less balanced distribution, with 175 memes in the training set (26.6%), 22 in the development set (32.8%) and 16 in the test set. The distribution of all classes, both binary and fine-grained, is illustrated in Figure 3.5.

To ensure a fair comparison between the models and to minimize the influence of the size of the dataset on the performance, the training datasets were reduced to similar sizes. Therefore, the same process was applied to the MAMI dataset, which was reduced to match the size of the EXIST 2024 dataset as allowed by the label distribution after removing non-overlapping instances and those without fine-grained labels. This resulted in 667 non-misogynous and 649 misogynous memes, totalling 1,316 which was close to the size of the EXIST dataset with 1,392 memes. For the fine-grained classes, the distribution closely resembled that of the EXIST dataset (with 1,053 before splitting into the final sets for cross-dataset experiments), with 449 instances in the stereotype class, 412 in objectification, and 197 in the violence category, with a total of 1,058 memes across all categories. After removing the non-overlapping classes, the MAMI dataset was also split into training and validation sets using the same ratio used for EXIST (90% and 10%, respectively). The statistics for the resulting splits are shown in Table 3.11 and 3.12. The distribution of all classes across the splits is presented in Figure 3.6.

Binary Label	**T$_o$**	**Train**		**Dev**		**Test (T$_o$)**	
		#	%	#	%	#	%
Misogynous	649	583	49	66	50	476	49
Non-misogynous	667	601	51	66	50	500	51
Total	1,316	1,184	100	132	100	976	100
Total T$_o$	2,292						

Table 3.11: MAMI binary class distribution after stratified sampling for cross-dataset experiments. T$_o$: Total with overlapping classes.

Fine-grained Label	**T$_o$**	**Train**		**Dev**		**Test (T$_o$)**	
		#	%	#	%	#	%
Objectification	412	369	63.3	43	65.2	348	73.1
Violence	197	176	30.2	21	31.8	153	32.1
Stereotype	449	402	69.0	47	71.2	350	73.5
Total	1,058	947	-	111	-	851	-
Total T$_o$	1,904						

Table 3.12: MAMI fine-grained class distribution after stratified sampling for cross-dataset experiments. T$_o$: Total with overlapping classes.

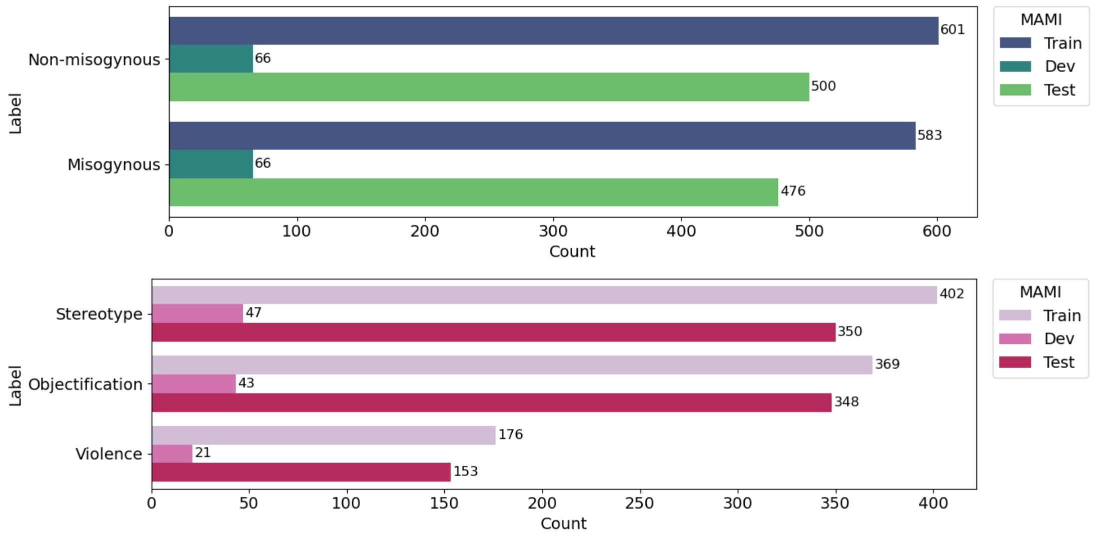


Figure 3.6: MAMI dataset splits after stratified sampling for cross-dataset experimentss. Distribution of binary (top) and fine-grained classes when meme is misogynous (bottom).

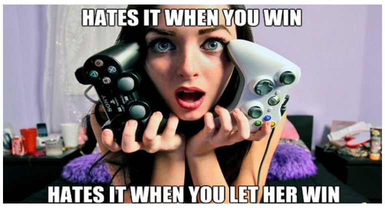
Meme	Meme text	Image caption
	HATES IT WHEN YOU WIN HATES IT WHEN YOU LET HER WIN	a girl holding up two video game controllers

Table 3.13: Meme, meme text and image caption from the BLIP-2 model.

3.2 Preprocessing

To implement the models that relied on textual information from the memes, it was necessary to obtain a textual representation of the meme images. This was achieved by extracting image captions using the BLIP-2 vision-language model, specifically the version incorporating FlanT5$_{\text{XL}}$ fine-tuned on COCO (Li et al., 2023). The initial prompt used was *"ignore text on the image. a photo of "*. However, upon inspecting the resulting image captions, many included text from the meme itself, either in addition to or instead of a description of the image. Therefore, phrases indicating the presence of meme text, such as "with the words" and "with the caption", were removed together with any subsequent tokens. In total, 3,982 memes from both datasets contained phrases that signalled the presence of meme text, out of which 3,220 belonged to the MAMI dataset and 642 to the EXIST dataset. The most common phrase was "with a caption", which appeared in a total of 1,952 memes in the MAMI dataset and 371 memes in the EXIST dataset. An overview of all such phrases and their frequency in each dataset is presented in Appendix A.1.

After this cleaning process, captions were checked to determine whether they were empty, included the word "meme", or contained at least three consecutive characters found in the meme text. This led to 110 memes requiring a second round of image captioning: 94 from the MAMI dataset and 16 from the EXIST 2024 dataset. For these memes, the image captioning process was repeated using the simpler prompt *"a photo of "*. The same filtering steps were applied to remove phrases that indicated the presence of text, resulting in the final dataset with image captions. It is worth noting that an initial trial using the simpler prompt on the full dataset did not improve the quality of the image captions; many were overly simplistic and repeated the meme text, confirming that the approach implemented was a better option. An example of a meme taken from the EXIST 2024 training dataset (ID: 210568) and its image caption is presented in Table 3.13.

The meme text provided in the dataset and image caption obtained with BLIP-2 were combined, creating a slightly different representation for each model processing only text. The meme text and image caption were concatenated with a full stop for the SVM model, with the special token [SEP] for BERT, and with the special token </s> for RoBERTa. This textual representation was lowercased before fine-tuning the RoBERTa model. While other preprocessing steps such as removing user mentions and urls were intially implemented, they did not provide an improvement in performance and were therefore disregarded.

3.3 Experimental Setup & Models

The experiments consisted of two setups: **in-domain** and **cross-dataset**. In the in-domain setup, each dataset was used to independently train and evaluate the models. In the cross-dataset setup, one dataset was used for training while the other was used to test the ability of the models to generalize to unseen data in a similar domain.

Following the approaches used in the EXIST 2024 and MAMI Shared Tasks, the experiments comprised (a) a **binary classification** task to distinguish between sexist or misogynous memes and non-sexist or non-misogynous memes, and (b) a **multi-label classification** task for which two different approaches were initially considered: hierarchical and flat. In the hierarchical setup, only memes identified as sexist or misogynous in step one were further classified into fine-grained classes in step two. In contrast, all labels were predicted simultaneously in the flat approach, i.e. both the sexist/misogynous and the fine-grained classes. Since each meme could be assigned to one or more fine-grained categories, both approaches in the multi-label classification task were implemented with a binary relevance strategy. This approach decomposes the multi-label problem into independent binary classifiers for each class label, ignoring the rest of the classes (Zhang et al., 2018). The multi-label classification strategy implemented in this thesis was BinaryRelevance[2] from the skmultilearn library (Szymański and Kajdanowicz, 2017).

Given that EXIST 2024 dataset was smaller, it served as the starting point for all experiments, as it was easier to test and refine the component models that would later be used in the subsequent experiments with the MAMI dataset in the in-domain setup, as well as in cross-dataset experiments. After experimenting with the flat and hierarchical approaches, it was determined that the hierarchical approach worked best for the multi-label classification task; therefore, the final experiments considered across all setups were implemented using this approach to predict the fine-grained classes.

The models used in the ensemble strategy included both text-only and and a state-of-the-art multimodal model. The text-only approaches consisted of a conventional SVM classifier and a pre-trained language model, while the multimodal model combined representations from a Swin Transformer-based visual model and a pre-trained language model. The multimodal approach also served as a **baseline** to assess whether the ensemble strategy that combined all these models could outperform it. All models were first tested on the binary classification task with the EXIST 2024 dataset to determine which would be the component models in the final ensemble. The component models and the hard majority voting ensemble strategy are described below.

3.3.1 Text-only Approaches

The text-only approaches implemented in this thesis, which relied on the meme text and image captions as input, included a conventional SVM classifier and a pre-trained language model. For the latter, BERT, RoBERTa and a BERTweet fine-tuned for sexism classification were explored, with RoBERTa demonstrating the best performance on the EXIST 2024 test set for in-domain binary classification, outperforming the other models by 0.1 and 0.2 points in macro F1 and ICM scores, respectively. As a result, RoBERTa was selected as the pre-trained component model across all experiments.

[2] http://scikit.ml/api/skmultilearn.problem_transform.br.html#binary-relevance

The SVM model with stylometric and emotion-based features as well as the pre-trained language models are described below.

Conventional Machine Learning: SVM

The SVM approach applied was selected since it has proven to effectively reduce the false positive rate when combined with transformer models in an ensemble (Markov and Daelemans, 2021), and has shown good cross-domain performance (Markov et al., 2021; Markov and Daelemans, 2021). Moreover, it has also produced uncorrelated predictions in comparison to pre-trained models when incorporated in an ensemble (Markov and Daelemans, 2021). This thesis implemented an SVM with stylometric and emotion-based features (Markov et al., 2021), which incorporated POS tags, function words, and emotion-conveying words and their associations from the NRC emotion lexicon (Mohammad and Turney, 2013).

The feature implementation was based on Markov et al. (2021) as follows. The first step was to lemmatize the textual representation of the memes, i.e. the meme text and image captions, with the NLTK library (Bird et al., 2009). Since POS tags can be indicators of harmful content by capturing morpho-syntactic patterns in text, universal POS tags (de Marneffe et al., 2021) were extracted with the same library. Function words help establish relationships between content-words in a sentence and introduce syntactic structures such as complements to verbs and relative clauses. Due to their role, they are considered the most significant type of stylometric features. In the context of emotion features, they can also serve as quantifiers or intensifiers (Markov et al., 2021). Therefore, the lemma of the function words were incorporated into the POS representation, which were identified with a set of closed class words.[3] Emotion information was incorporated with the NRC emotion lexicon (Mohammad and Turney, 2013), which contains 6,468 unique English terms and their associations with eight emotion words–anger, anticipation, disgust, fear, joy, sadness, surprise, and trust–as well as with positive or negative sentiment, i.e how strongly the term is associated with the emotion. The emotion-conveying words were added to the POS and function words representation. In addition, they were represented as features on their own by keeping the emotions and the sentiment that were associated to the textual representation of the memes.

For example, the textual representation of meme ID 210332 from the EXIST 2024 dataset was "me Listening to the same sad song on repeat, making sure it does enough damage. a woman in a bathtub with a heart shaped heart in the middle", which was represented through POS, function words, and emotion-conveying words as "me VERB PRT the ADJ NOUN NOUN on NOUN . VERB ADJ it VERB ADJ damage . a NOUN in a NOUN with a NOUN VERB NOUN in the NOUN". The emotion association feature from the NRC emotion lexicon of this meme was "anger disgust negative sadness".

To determine the best feature configuration for the SVM approach, an ablation study was conducted. Table 3.14 presents the most significant setups explored to identify the configuration that yielded the best performance on the EXIST development split that would also generalize to the MAMI dataset in the binary classification task. Incorporating surface-level textual representation improved the macro-F1 score. While processing POS, function words and emotion-conveying words as unigrams, bigrams

[3] https://universaldependencies.org/u/pos/#universal-pos-tags

POS + FW + Emo			Emotion Associations			Textual Representation				Results		
WS	Analyser	n-gram	WS	Analyser	n-gram	WS	Analyser	n-gram	min df	Non-sexist	Sexist	Macro F1
TF-IDF	Word	1-1	TF-IDF	Word	2-2	TF-IDF	Character	1-6	7 15	0.46 0.52	0.67 0.68	0.56 0.60
						TF-IDF	Character	3-6	7 10 15 20	0.50 0.50 0.59 0.53	0.68 0.67 0.68 0.69	0.59 0.59 0.64 0.61
						TF + TF-IDF	Character	3-6	-	0.62	0.69	**0.66**
								3-6 1-2	-	0.62	0.69	**0.66**
						TF	Character	1-6	7 10	0.63 0.59	0.68 0.64	**0.66** 0.61
								2-6	-	0.62	0.68	0.65
								3-7	-	0.59	0.68	0.64
								3-6	7 -	0.61 0.62	0.66 0.69	0.63 **0.66**
			TF-IDF	Word	2-3	TF	Character	3-6	-	0.62	0.69	**0.66**
								1-6	7	**0.64**	0.69	**0.66**
			TF-IDF	Word	1-2	TF	Character	3-6	-	0.60	0.68	0.64
TF-IDF	Word	1-3	TF-IDF	Word	1-1	TF-IDF	Character	3-7	-	0.53	**0.70**	0.62
							Character + Word	3-7 1-3	-	0.53	**0.70**	0.61
							Word	1-2	-	0.51	0.69	0.60
						TF	Character	4-7	-	0.53	**0.70**	0.62
							Character + Word	4-7 1-3	-	0.60	0.67	0.64
						-	-	-	-	0.40	0.64	0.52

Table 3.14: Performance of different parameter configurations across features implemented in SVM approach on EXIST 2024 dataset. POS: part-of-speech; FW: function words; Emo: emotion conveying words; WS: weighting scheme. The selected configuration for the SVM approach is shown in pink.

and trigrams enhanced the performance for the sexist class, in particular when combined with character-level surface features, it did not outperform using unigrams alone in terms of macro-F1 score. Moreover, even though the non-sexist class showed better performance when bigrams and trigrams were used for the emotion associations feature with a TF representation of the surface-level features, this configuration did not generalize well to the MAMI dataset. In contrast, using only bigrams for emotion associations performed better across datasets. The same occurred when experimenting with different n-gram ranges for the texual features. Including both the TF and TF-IDF representations for texual representation did not improve performance, nor did applying a minimum occurrence threshold; both were therefore disregarded. This pattern was the case for multi-label classification as well. Therefore, the final configuration selected involved vectorizing POS, function words and emotion-conveying words and their emotion and sentiment associations from the NRC emotion lexicon using a TF-IDF weighting scheme. The former were extracted as unigrams, while emotion associations were extracted as bigrams. Character n-grams (with n=3–6) were extracted from the meme text and image captions using a TF weighting scheme. All these features were combined in the final SVM setup.

The model was built with the liblinear implementation of SVM from scikit-learn (Pedregosa et al., 2011). The regularization parameter (C) was optimized through a grid search, with the random state set to 0 for reproducibility and the maximum number of iterations set to 30,000 due to the high dimensionality of the vector created from the MAMI dataset. Due to the size of this dataset, the inclusion of additional parameters, such as tolerance for stopping criteria and the loss function, was impractical given that doing so required more than 24 hours to process both binary and multi-label classification tasks for this dataset only. The values explored for the C parameter were 0.1, 0.5, 1, 5, 10 to control the trade-off between the training error and the margin width, both in binary classification and multi-label classification. The best value for C was 0.1 throughout the datasets and classification tasks, both in in-domain and cross-dataset settings.

Pre-trained Language Model

The second text-only approach implement as a component model was a pre-trained language model due to the improvement they brought to the detection of harmful content. In order to determine what would work best in the ensemble, different models were explored, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and a BERTweet model fine-tuned for sexism detection (Al-Azzawi et al., 2023). These models are based on the transformer architecture (Vaswani et al., 2017), a neural network with an encoder-decoder structure composed of three main components: an input encoding, transformer blocks, and a final linear transformation followeb by a softmax layer for output prediction. A brief overview of the transformer architecture is presented below.

Each token in the input sequence is first embedded with both token and positional embeddings, which are then passed through multiple transformer layers. Each transformer block includes a multi-head attention mechanism and a feedforward neural network, both preceded by layer normalization and connected via residual connections. The multi-head attention mechanism builds contextual embeddings by comparing each token with surrounding tokens using query, key, and value vectors: the query represents the current token being compared to others; the key represents other tokens used to determine similarity weights; and the value provides the content to be weighted and summed up to produce the output for the current token. This mechanism enables the model to capture various relational aspects between words by incorporating contextual information (e.g., disambiguating “bank” depending on whether words like “river” or “money” appear nearby) and learning complex mappings between inputs and outputs. The encoder processes the full sequence with bidirectional attention, allowing the model to attend to all tokens, both before and after the current one. In contrast, the decoder uses causal attention, where each token attends only to preceding tokens. After passing through all layers, the output of the decoder is used to predict the next token via a linear transformation followed by softmax function (Jurafsky and Martin, 2025).

Based on the above explanation, the language models implemented are encoder-only models, also known as bidirectional encoders. These models are used to produce high-quality embeddings for text classification tasks (Jurafsky and Martin, 2025), as is the case of hate speech detection. The first model considered was **BERT** (Devlin et al., 2019). This model processes single sentences or sentence pairs combined into a sequence, adding the special [CLS] token at the beginning and the [SEP] to separate sentences. Segment embeddings indicate whether a token belongs to the first or the second sentence. BERT uses WordPiece subword tokenization with a vocabulary size of 30,000 tokens and combines token, segment, and position embeddings before passing the input into transformer blocks. With a combined total of 3.3 billion words, BERT was pre-trained on unlabelled data from BooksCorpus and English Wikipedia with two objectives: masked language modelling (MLM), where 15% of tokens in the input were masked and predicted using both the left and right context, and next sentence prediction (NSP), where it predicted whether pairs of sentences were adjacent or unrelated. These tasks enabled BERT to learn contextual word representations (Devlin et al., 2019; Jurafsky and Martin, 2025). The BERT$_{\text{BASE}}$[4] variant of the model was selected for the experiments, in particular the model that does not differentiate between uppercased and lowercased characters (uncased). It features 12 transformer blocks, 768 hidden units, 12 attention heads, 110M parameters, and supports a maximum input sequence length of 512 tokens.

The second approach was **RoBERta** (Liu et al., 2019), which stands for Robustly optimized BERT approach. It is a BERT-based model that showed improved performance over the original BERT (Devlin et al., 2019) by training for longer using bigger batches and more data, specifically, five English corpora: BookCorpus, English Wikipedia, CC-News, OpenWebText, and Stories. RoBERTa employs BPE (Byte-Pair Encoding) to split training data into subword units by selecting the most frequently occurring symbol pairs. It removed the NSP objective from BERT and was trained on longer sequences. In addition, it dynamically altered the masking pattern applied to the training data for the MLM objective (Liu et al., 2019). RoBERTa$_{\text{BASE}}$[5] was implemented in this thesis. Like the BERT$_{\text{BASE}}$ model, this version of RoBERTa has 12 transformer blocks, 768 hidden units, 12 attention heads, and supports sequences of up to 512 tokens; however, it contains more parameters than BERT, with a total of 125M.

The third approach was a **BERTweet** model, which is a large-scale pre-trained language model for English Tweets. While it has the same architecture as BERT$_{\text{BASE}}$, this model was trained using the pre-training procedure from RoBERTa. The training data consisted of 873M English tweets wight a length of 10 to 64 tokens (Nguyen et al., 2020). The version with a large architecture has 355M parameters and supports a maximum length of 512 tokens.[6] The model implemented was BERTweet$_{\text{LARGE}}$ fine-tuned for sexism detection[7] (Al-Azzawi et al., 2023) using the EDOS dataset (Kirk et al., 2023).

The application of these models for sexism and misogyny identification and categorization was achieved through **fine-tuning**, which involves building an application-specific head on top of the pre-trained model that takes its output as input. Task-specific parameters are added, and the model is further trained with gold-labelled supervised data, in this case for the detection of sexism or misogyny. During this process, the parameters of the pre-trained model are either frozen or minimally updated. Fine-tuning pre-trained models is a form of transfer learning, a method where knowledge acquired from one task or domain is applied to solve a new task (Jurafsky and Martin, 2025).

As mentioned above, the experimental setup was first tested on the EXIST 2024 dataset. Therefore, these pre-trained models were fine-tuned for sexism identification to determine which model to implement across all experimental setups. As shown on Table 3.15, while BERTweet fine-tuned for sexism detection performed best on

[4] https://huggingface.co/google-bert/bert-base-uncased

[5] https://huggingface.co/FacebookAI/roberta-base

[6] https://github.com/VinAIResearch/BERTweet

[7] https://huggingface.co/NLP-LTU/bertweet-large-sexism-detector

EXIST 2024	**Dev**					**Test**				
Model	F1					F1				
	Macro	Micro	Pos. Class	ICM Norm	Mean	Macro	Micro	Pos. Class	ICM Norm	Mean
BERT	0.73	0.73	**0.76**	0.59	0.70	0.70	0.71	0.75	0.55	0.68
RoBERTa	0.73	0.73	**0.76**	0.59	0.70	**0.71**	**0.72**	**0.76**	**0.57**	**0.69**
BERTweet$_{\text{LARGE}}$ sexism-detector	**0.75**	**0.75**	**0.76**	**0.63**	**0.72**	0.70	0.70	0.73	0.55	0.67

Table 3.15: Performance of explored pre-trained models for sexism identification on EXIST 2024 dataset.

the development set, RoBERTa demonstrated stronger performance across all metrics implemented (which will be described in the following section) on the test set. This might be due to the BERTweet model overfitting to the data. As a result, **RoBERTa** was selected as the pre-trained component model to be incorporated in the ensemble across all tasks performed in this thesis. Its better performance might be related to the larger datasets used during its pre-training and the larger vocabulary size resulting from BPE, compared to the other BERT-based models.

Throughout the experiments with pre-trained models, hyper-parameter optimization was performed using Optuna (Akiba et al., 2019), which parallelises and iterates through predefined ranges to find the optimal configuration for a model. The parameters explored for binary and multi-label classification in both in-domain and cross-dataset settings were batch size, learning rate, weight decay, and number of epochs. The following values were considered:

- Batch size: 8, 16, 32, and 64, to identify the optimal number of training samples processed before model update.

- Learning rate: Ranged from 5e-6 to 1e-4, to determine the step size used by the optimization algorithm when updating the weights of the model to minimize the loss function.

- Weight decay: Ranged from 1e-6 to 0.01. This parameter adds a regularization penalty to the loss function that increases with the magnitude of model weights, helping to prevent overfitting by keeping the weights smaller.

- Number of epochs: Between 2 and 10, to determine the amount of times the model gets exposed to the entire training dataset during training.

The optimal set of parameters was determined for each classification task, and in the case of multi-label classification, for each label since binary relevance was used. The resulting parameters used to fine-tune RoBERTa are presented in Table 3.16. The experiments were conducted on the Google Colaboratory platform with an NVIDIA A100 GPU. The Transformers library[8] (Wolf et al., 2020) form the Hugging Face platform was used, and a random seed was set for reproducibility.

[8] https://github.com/huggingface/transformers

	Dataset	Label	Batch Size	Learning Rate	Weight Decay	Epochs
In-Domain	EXIST 2024	Sexist	64	3,70E+10	1,41E+10	7
		Ideological and inequality	64	7,59E+09	1,48E-03	10
		Misogyny and non-sexual violence	32	8,07E+09	7,06E+10	10
		Objectification	64	2,49E+11	4,82E+09	10
		Sexual violence	64	1,45E+11	4,95E-03	10
		Stereotyping and dominance	8	2,37E+09	3,02E+10	6
	MAMI	Misogynous	64	3,10E+11	2,78E-04	7
		Shaming	8	5,34E+09	4,26E+09	9
		Stereotype	32	7,01E+09	8,25E-03	9
		Objectification	64	1,61E+10	1,75E+10	8
		Violence	32	1,03E+11	4,71E+10	8
Cross-Dataset	EXIST 2024	Sexist	64	1.35E+11	7.80E-04	10
		Objectification	64	2.43E+11	1.55E+09	6
		Sexual violence	32	1.88E+10	5.36E+09	8
		Stereotyping and dominance	8	1.04E+11	1.10E-04	7
	MAMI	Misogynous	16	2.94E+11	5.92E+09	6
		Objectification	8	1.26E+11	7.10E+10	8
		Violence	64	3.29E+10	6.84E+09	10
		Stereotype	16	1.49E+10	2.60E+10	6

Table 3.16: Best parameters found with Optuna to fine-tune RoBERTa for each task.

3.3.2 Multimodal Approach

The multimodal approach was implemented following the architecture introduced by Wang and Markov (2024a,b,c), which, as explained in Chapter 2, achieved state-of-the-art in the classification of harmful memes. In this system, a vision model is used to extract embeddings from the meme image and a language model is used to extract textual embeddings from the meme text. The resulting visual and textual representations are then concatenated and passed through an MLP fusion module (Shi et al., 2021), followed by a prediction layer to classify each instance. Following the process used to select the language model for the text-only approach, different models and combinations were explored to determine the best multimodal component model for the ensemble with BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) for extracting textual embeddings, and ViT (Dosovitskiy et al., 2021) and Swin Transformer V2 (Liu et al., 2022) for visual features.

To understand how a vision model works, I will first briefly introduce image processing. While the most basic unit of information in NLP is a word or token, in computer vision it is an image. Thus, similar to how natural language is vectorized, images must also be converted into numerical representations for machines to process them. The pixels of an image, which range from hundreds to thousands, contain information about colour and light intensity. In greyscale images, the intensity of each pixel can be represented by a value ranging from 0 to 255, indicating its level of darkness. In contrast, coloured images are usually stored in the RGB (Red, Green, Blue) system, where each pixel consists of three separate values corresponding to the intensity of the red, green, and blue channels. In such case, an image is represented by three channels, with the combination of values determining the colour of each pixel.[9]

Dosovitskiy et al. (2021) introduced the **Vision Transformer (ViT)**, which applies the encoder block of the transformer architecture for image classification. ViT

[9] `https://www.datacamp.com/blog/what-is-computer-vision` accessed on 27th May 2025.

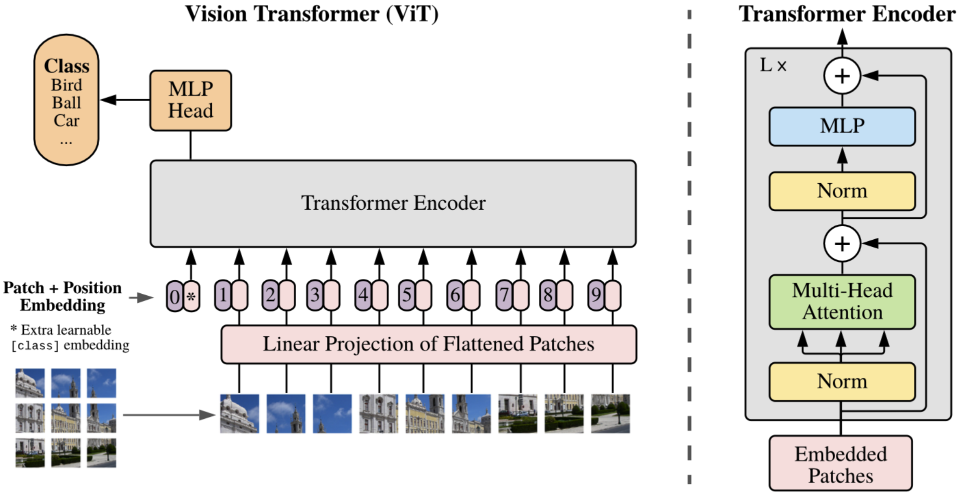


Figure 3.7: ViT architecture (taken from Dosovitskiy et al. (2021))

processes images by splitting them into patches, which are flattened and linearly projected into fixed-size vectors known as patch embeddings. A special learnable token is prepended to the sequence of patch embeddings to represent the entire image, and positional embeddings are added to retain positional information. This sequence is then passed through the transformer encoder, and the output corresponding to the classification token is used as the overall image representation. Based on this representation, a classification head is used to make predictions, which is implemented as an MLP during pre-training and a single linear layer during fine-tuning. Figure 3.7 provides an overview of the model as illustrated by the authors. ViT was pre-trained and fine-tuned on large-scale image datasets including ImageNet (1k and 21k classes) and JFT (with 18K classes). The version of ViT implemented to explore the multimodal model combinations has 102.6M parameters and processes images at a resolution of 224x224, which are presented to the model as a sequence of fixed-size patches with a resolution of 16x16.[10]

Since image resolutions can range from low to high, processing high-resolution images with a transformer becomes increasingly hard, as the computational complexity of self-attention grows quadratically with image size. To address this, Swin Transformer (Liu et al., 2021) builds a hierarchical representation by starting with small image patches and gradually merging neighbouring patches in deeper layers. In this design, self-attention is computed locally within non-overlapping windows that partition the image. Because each window contains a fixed number of patches, the computational complexity becomes linear with respect to image size. This model applies several transformer blocks that modify the standard self-attention mechanism by replacing multi-head self attention with a module based on regular and shifted window partitioning configurations. This means that a patch might be grouped with different neighbouring patches across layers, enabling better information flow across different regions of the image and resulting in more powerful and connected image representations. The shifted window processed is illustrated in Figure 3.8(a), as opposed to how ViT processes patches in (b). **Swin Transformer V2** can process high-resolution images and

[10] https://huggingface.co/google/vit-base-patch16-224

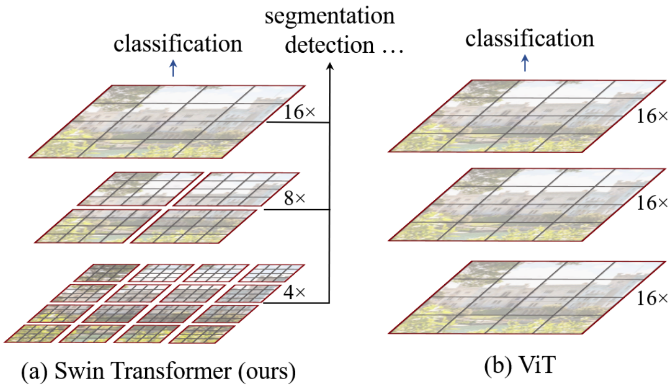


Figure 3.8: A comparison of Swin and ViT feature mapping (taken from Liu et al. (2021))

EXIST 2024	Dev	Test				
	F1	F1				
Multimodal Model	Macro	Macro	Micro	Pos. Class	ICM Norm	Mean
BERT+Swin Transformer V2	0.72	0.67	0.68	0.72	0.51	0.64
RoBERTa+Swin Transformer V2	**0.76**	**0.73**	**0.74**	**0.76**	**0.60**	**0.71**
BERT+ViT	0.73	0.67	0.68	0.72	0.51	0.64

Table 3.17: Performance of the multimodal models explored for sexism identification on EXIST 2024 dataset.

employs a self-supervised pre-training method to reduce reliance on large amounts of labelled data (Liu et al., 2022). The base version of the Swin Transformer V2 was used, which has 88M parameters and processes 256 x 256 resolution images with a window size of 8.[11]

To implement the multimodal approach, the Swin Transformer V2 and ViT models mentioned above were used to extract visual features, while the RoBERTa and BERT models described in the text-only approaches were used to extract contextualized textual embeddings during experiments on the EXIST 2024 dataset. Following the implementation from Wang and Markov (2024a,b,c), the textual and visual representations were concatenated and passed through an MLP fusion module (Shi et al., 2021), followed by a prediction layer for classification. Swin Transformer V2 was combined with both BERT and RoBERTa, with the combination involving RoBERTa yielding better performance. To explore another vision model, ViT was implemented in combination with BERT, but it did not outperform the combination of **Swin Transformer V2 and RoBERTa**, as shown in Table 3.17. This might be attributed to the powerful image representation Swin Transformer achieves through its hierarchical architecture and representations computed through shifted windows, as well as the strong textual representations provided by RoBERTa, which benefits from pre-training on larger datasets, an advantage that was already confirmed in the text-only approach above. This model also served as a multimodal baseline throughout the experiments.

As with the pre-trained language models used for the textual modality, the experiments for the multimodal approach were also conducted on the Google Colaboratory

[11] https://huggingface.co/microsoft/swinv2-base-patch4-window8-256

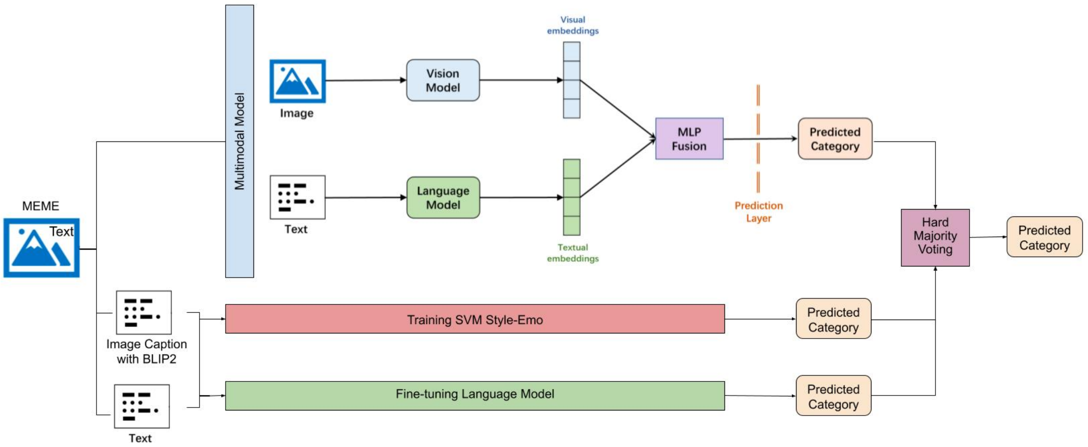


Figure 3.9: Overall architecture of implemented ensemble approach. The multimodal architecture (upper part of the figure) is based on Wang and Markov (2024c).

platform using an NVIDIA A100 GPU. The PyTorch framework was employed, with the timm[12] (PyTorch Image Models) library for the vision models and the Transformers library (Wolf et al., 2020) for the language models. Deterministic algorithms were used, and a fixed random seed was set to ensure reproducibility. The experiments were run with consistent hyperparameters: a base learning rate of 1e-5, a batch size of 16, a maximum of 10 training epochs, and optimization using the AdamW optimizer[13] with cross-entropy loss.

3.3.3 Ensemble Strategy

An ensemble combines multiple models to generate a final prediction, based on the premise that the errors of individual models can offset one another, resulting in improved overall predictive performance compared to relying on a single model (Sagi and Rokach, 2018). The ensemble strategy implemented in this thesis was majority voting, where the goal is to output the most common prediction among the individual models. Specifically, hard majority voting was used, in which each model provides a discrete prediction, and the final output is determined by the label receiving the most votes. This approach treats all models as equally reliable without considering their confidence in their predictions (Kyriakides and Margaritis, 2019).

Apart from the SVM with stylometric and emotion-based features, the best component models demonstrated to be the RoBERTa, and the multimodal Swin Transformer V2 with RoBERTa. The the ensemble strategy combined their predictions to output a final prediction for each instance. Figure 3.9 provides an overview of the system implemented, which was applied to both binary and multi-label classification tasks. In the latter, the ensemble handled cases where no fine-grained labels were predicted despite a positive prediction at the binary level by converting the prediction to the negative class at the binary level.

[12] https://github.com/huggingface/pytorch-image-models

[13] https://docs.pytorch.org/docs/stable/generated/torch.optim.AdamW.html

If the ensemble did not achieve better or comparable performance to the individual component models, the best component model was used to generate the predictions of the first step in multi-label hierarchical classification.

3.4 Evaluation Metrics

The evaluation metrics used were based on those defined in the shared tasks from which the datasets originated. The EXIST 2024 Shared Task (Plaza et al., 2024b) implemented ICM (Information Contrast Measure) metric (Amigó and Delgado, 2022) and a normalized version of ICM (ICM Norm) to evaluate both binary and multi-label classification. In addition, they included the F1 of the positive class, i.e., the sexist category, for binary classification, and macro F1 for multi-label classification. In contrast, the MAMI Shared Task evaluated binary classification with macro F1 and multi-label classificiation with weighted F1. Apart from these metrics, micro F1 was implemented in this thesis, as it is a commonly used metric in harmful content detection. Since the metrics varied across the shared tasks, the mean of all of them was calculated to better reflect the overall performance of the models. The metrics are described below.

ICM Norm ICM is a similarity function that extends Pointwise Mutual Information (PMI) and is used to evaluate system outputs in classification tasks by measuring their similarity to the ground truth categories. The general formulation of ICM is suited for scenarios where categories have a hierarchical structure and items can belong to multiple categories. It is calculated with the below formula:

$$ICM(s(d), g(d)) = 2I(s(d)) + 2I(g(d)) - 3I(s(d) \cup g(d))$$

Here, $I()$ denotes Information Content, $s(d)$ represents the set of categories assigned to document d by the system s, and $g(d)$ refers to the set of categories assigned to document d in the gold standard, as explained by the organizers of the shared task in the guidelines for the participants. Given that ICM output values in a range of -1 to 1, the ICM Norm was considered for the experiments in this thesis.

F1-score $F1$ evaluates the performance of a model by computing the harmonic mean of precision and recall. It is the ability of a model to accurately identify the cases where it predicted the category (P), and to identify positive cases of it (R) (Jurafsky and Martin, 2025). It is defined as:

$$F1 = \frac{2PR}{P + R}$$

The following F1-scores were computed:

- **F1 of the positive class**, for binary classification.
- **Macro F1**, for both binary and multi-label classification. This metric averages the individual F1 scores of each class, regardless of the number of samples in them.

- **Micro F1**, for both tasks as well. It measures the proportion of correctly classified observations out of all observations.

  In binary classification, micro F1 corresponds to accuracy, which is defined as the percentage of correct classifications, calculated as $1 - error\ rate$. The formula to calculate accuracy is $Accuracy = \frac{TP+TN}{All\ Samples}$ True positives (TP) are the true correct answers while true negatives (TN) are the true wrong answers. In the tasks at hand, true positives are memes that the model predicted as positive cases (sexist or misogynous, for instance) and they were indeed labelled as such. On the contrary, true negatives are those that were labelled as negative (non-sexist or non-misogynous) in the gold-labelled memes, and were also predicted as negative by the model.

  In multi-label classification, it is calculated by summing the TP, false negatives (FN) and false positives (FP) across all classes, and then computing the F1 score based on these aggregated values.[14]

- **Weighted F1**, for multi-label classification. It is a weighted average of the F1 scores for each class, with weights based on the number of samples in each of them. In the MAMI Shared Task, this metric was computed across the fine-grained classes, ignoring the misogynous class. The same was implemented in the experiments conducted.

Mean The **average** of all the metrics for each task was computed to enable a comprehensive comparison of the performance of the models. In binary classification, the mean averaged ICM Norm, F1 of the positive class, macro F1 and micro F1; while it averaged ICM Norm, macro F1, weighted F1, and micro F1 scores for multi-label classification.

The metrics that originated in the EXIST 2024 Shared Task were implemented with the PyEvALL library,[15] i.e. ICM Norm, F1 of the positive class and macro F1 for multi-label classification. The macro F1 for binary classification and weighted F1 scores were computed with the functions provided by the organizers of the MAMI Shared Task.[16] Furthermore, the micro F1 scores was retrieved from the classification report from scikit-learn (Pedregosa et al., 2011).

To summarise, the metrics used for binary classification were ICM Norm, F1-score of the positive class, macro and micro F1, and the average of all these metrics. For multi-label classification, ICM Norm, macro F1, weighted F1, and micro F1 scores, and their average were used instead.

[14] https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f/

[15] https://github.com/UNEDLENAR/PyEvALL

[16] https://github.com/MIND-Lab/SemEval2022-Task-5-Multimedia-Automatic-Misogyny-Identification-MAMI-/tree/main

Chapter 4

Results and Analyses

This chapter describes the results obtained in a quantitative and qualitative manner. It starts with a quantitative analysis that discusses the performance of the models in binary and multi-label classification in both in-domain and cross-dataset setups. A qualitative analysis follows with a correlation and detailed error analysis with the aim of investigating why ensembles are helpful for the identification and categorization of sexism and misogyny in the evaluated settings, as well as identifying their limitations.

4.1 Quantitative Analysis

The results are presented per experimental setup, beginning with the performance achieved in the binary classification task, followed by multi-label classification task. Within each task, the results are described separately for in-domain and cross-dataset setups to address the research questions. It is important to highlight that the state-of-the-art results on the EXIST 2024 dataset were obtained using the original test set, which was not publicly available. Therefore, the results presented in this research are not directly comparable, as the test set used here was derived by splitting the original meme training set.

4.1.1 Binary Classification

The performance of the models for binary classification across evaluation settings is presented in Table 4.1. The in-domain setting refers to the experiments in which training and testing were conducted independently on the EXIST 2024 and MAMI datasets, while training was performed with one dataset and test with the other in the cross-dataset experiments.

In-Domain

The results in the in-domain setting show that the ensemble strategy outperformed the component models on the EXIST dataset, but it did not outperform the individual models on the MAMI dataset, where the fine-tuned RoBERTa obtained the highest results.

For the EXIST dataset, the best component model was the multimodal baseline model that combined Swin Transformer V2 with RoBERTa, with a mean score of 0.71. All individual models were outperformed by the ensemble, which obtained the best performance across all metrics for sexism identification, with a mean score of 0.72.

In-Domain	**EXIST 2024**					**MAMI**				
	F1					F1				
Model	Macro	Micro	Pos. Class	ICM Norm	Mean	Macro	Micro	Pos. Class	ICM Norm	Mean
Baseline:RoBERTa+ Swin Transformer V2	0.73	0.74	0.76	0.60	0.71	0.67	0.69	0.75	0.54	0.66
SVM Style-Emo	0.64	0.65	0.71	0.47	0.62	0.65	0.66	0.72	0.50	0.63
RoBERTa	0.71	0.72	0.76	0.57	0.69	**0.71**	**0.72**	**0.76**	**0.58**	**0.69**
Ensemble	**0.74**	**0.75**	**0.78**	**0.61**	**0.72**	0.69	0.71	**0.76**	0.57	0.68
Cross-Dataset	**MAMI - EXIST 2024**					**EXIST 2024 - MAMI**				
	F1					F1				
Model	Macro	Micro	Pos. Class	ICM Norm	Mean	Macro	Micro	Pos. Class	ICM Norm	Mean
Baseline:RoBERTa+ Swin Transformer V2	0.65	0.65	0.67	0.48	0.61	0.68	0.68	**0.71**	**0.53**	**0.65**
SVM Style-Emo	0.62	0.62	0.60	0.43	0.57	0.60	0.60	0.63	0.41	0.56
RoBERTa	0.66	0.66	0.68	0.49	0.62	0.68	0.68	0.67	0.51	0.64
Ensemble	**0.67**	**0.67**	**0.69**	**0.51**	**0.64**	**0.69**	**0.69**	0.70	**0.53**	**0.65**

Table 4.1: In-domain and cross-dataset results for the multimodal baseline, component models, and ensemble in binary classification. Pos. Class: F1 score of positive class on each dataset.

Figure 4.1 presents the confusion matrices for component models and the ensemble on the EXIST 2024 dataset. The ensemble for sexism classification showed the fewest errors across models, with a total of 43 errors (19 FN and 24 FP).

On the MAMI dataset, the best component model for misogyny identification was the text-only RoBERTa approach, with a mean score of 0.69. While the ensemble strategy outperformed the multimodal baseline and the SVM approach, with a mean score of 0.68, it did not provide better results than RoBERTa. The ensemble obtained the same F1 score of 0.76 as RoBERTa for the misogynous class. Figure 4.2 presents the confusion matrices for component models and ensembles on the MAMI dataset. The text-only RoBERTa model made the lowest number of misclassifications. All models misclassified non-misogynous memes as misogynous ones, being more oriented towards recall than precision on the positive class. As noted by Fersini et al. (2022) on the MAMI shared task, this bias suggests that the models might be misled by the presence of certain type of text or images.

Cross-Dataset

Moving to the cross-dataset experiments, it can be observed that the ensemble model performed well across both datasets. It clearly outperformed the component models when the MAMI dataset was used for training and the EXIST 2024 dataset for testing in the identification of sexism, achieving a mean score of 0.64. In this setup, even though RoBERTa was the strongest individual model–as occurred in the in-domain experiments on the MAMI dataset–, the ensemble obtained the highest scores across all evaluation metrics.

A drop in performance was observed for all models when trained on misogyny and tested on sexism. The ensemble, while still the best performing system, showed a performance decline that ranged from 0.07 to 0.10 points. The multimodal baseline,

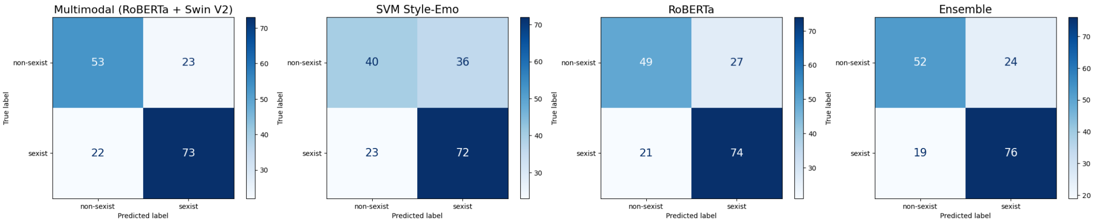


Figure 4.1: Confusion matrices for the component models and ensemble approach in binary classification on EXIST 2024.

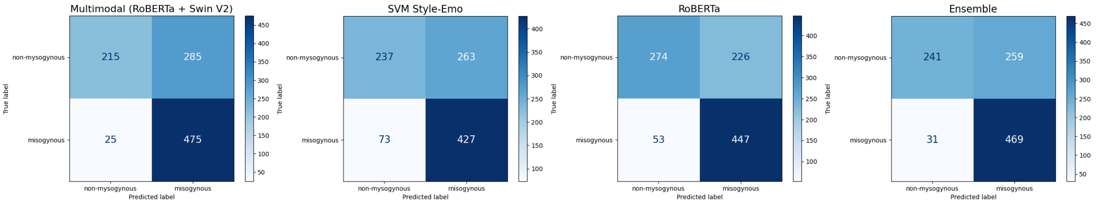


Figure 4.2: Confusion matrices for the component models and ensemble approach in binary classification on MAMI.

which was the best component model in the in-domain setting, experienced a 0.08 to 0.12-point drop across evaluation metrics. The text-only approaches also declined, with RoBERTa showing a 0.07 point decrease in the mean score and SVM only 0.05. While SVM experienced its largest drop on the sexist class (a decrease of 0.11 points), it exhibited the smallest performance drop in the mean score among all models when trained on MAMI and tested on EXIST 2024 for sexism identification in memes. This aligns with the findings of Markov et al. (2021) and Markov and Daelemans (2021), who reported that an SVM with stylometric and emotion-based features was beneficial in a cross-domain setting.

Figure 4.3 presents the confusion matrices for cross-dataset sexism identification, where the ensemble made 50 mistakes (20 FN and 30 FP), being the model with the lowest number of errors. Interestingly, while the ensemble is the system with the least amount of FN, the SVM model produced the fewest FP, with only 25 misclassifications. This model contributed to a lower FP rate in the identification of harmful content in text (Markov and Daelemans, 2021), and has now demonstrated to provide the same benefits for the identification of sexist memes in a cross-dataset setting.

When training on the EXIST 2024 dataset and testing on MAMI for misogyny identification, the multimodal baseline that combined Swin Transformer V2 with RoBERTa was the best component model, with a mean score of 0.65. This also occurred in the in-domain experiments for sexism identification. The ensemble strategy also achieved a mean score of 0.65. While the baseline model performed better on the misogynous class, with a F1 score of 0.71 vs. 0.70 from the ensemble, the ensemble outperformed all component models in macro and micro F1. This suggests that the ensemble was better at also capturing the non-misogynous memes.

The drop from in-domain to cross-dataset showed different patterns when testing on the MAMI dataset. The multimodal model dropped 0.04 points in the positive class,

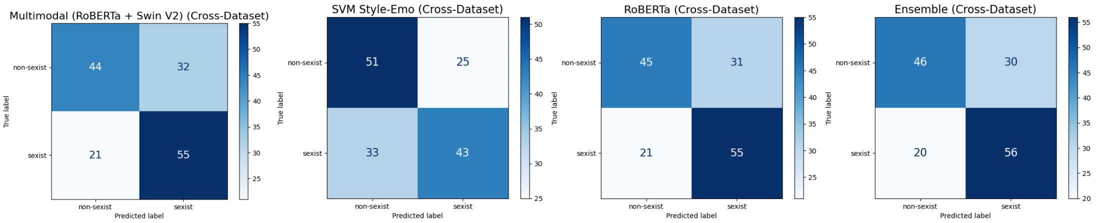


Figure 4.3: Confusion matrices for the component models and ensemble approach in binary classification when training on MAMI and testing on EXIST 2024.

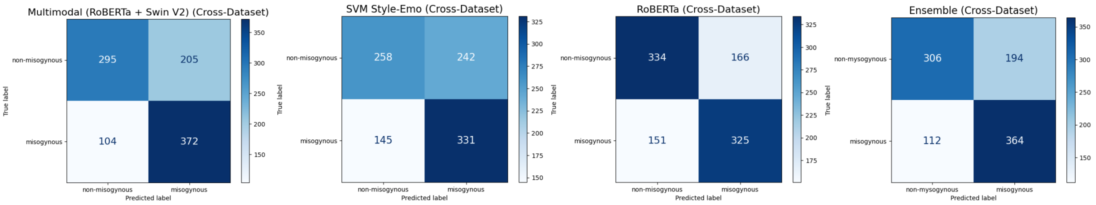


Figure 4.4: Confusion matrices for the component models and ensemble approach in binary classification when training on EXIST 2024 and testing on MAMI.

and 0.01 point across the rest of the evaluation metrics. The fine-tuned RoBERTa and the SVM model showed a drop of 0.05 to 0.09 points across all metrics. While the ensemble also declined in performance on most of the metrics (with 0.02 to 0.06 points drop), it achieved the same macro F1 in both in-domain and cross-dataset settings. The multimodal baseline showed the smallest performance drop. In general, the drop in this setting was rather small, considering that the dataset used for training was much smaller than in in-domain (1,252 vs. 9,000 memes).

Figure 4.4 presents the confusion matrices of the models for cross-dataset misogyny identification. The ensemble made the lowest number of errors, with 306 (112 FN and 194 FP), followed by the multimodal model, with 309 misclassified memes (104 FN and 205 FP). This helps explain why the ensemble achieved higher macro and micro F1 scores: it made fewer mistakes on the non-misogynous class but correctly identified fewer misogynous memes. The confusion matrices indicate that an ensemble trained on the smaller sexist meme dataset has a more balanced precision and recall than when training on the original MAMI dataset, despite making more errors on the positive class.

In general, the models showed similar performance on both datasets, achieving similar results across the evaluation metrics in the cross-dataset setup, with only 0.1 and 0.2 points difference affecting the performance on the EXIST dataset for sexism identification. Since the training datasets in this setup had similar sizes, these results indicate that they generalized well to each other, being similar phenomena.

Summary

Based on the results for binary classification in the in-domain setting, an ensemble strategy combining the multimodal Swin Transformer V2 with RoBERTa, a fine-tuned RoBERTa model, and SVM with stylometric and emotion-based features demonstrated the strongest performance in the identification of sexism in memes, outperforming all

In-Domain	EXIST 2024					MAMI				
	F1					F1				
Model	Macro	Micro	WA	ICM Norm	Mean	Macro	Micro	WA	ICM Norm	Mean
Baseline: RoBERTa+ Swin Transformer V2	0.48	0.56	0.62	0.35	0.50	**0.55**	**0.58**	**0.68**	**0.44**	**0.56**
SVM Style-Emo	0.46	0.56	**0.64**	0.33	0.49	0.51	0.55	0.65	0.40	0.53
RoBERTa	**0.50**	**0.57**	0.63	0.36	**0.52**	0.54	0.56	0.67	0.43	0.55
Ensemble	0.49	**0.57**	**0.64**	**0.37**	**0.52**	**0.55**	0.57	**0.68**	**0.44**	**0.56**

Cross-Dataset	MAMI - EXIST 2024					EXIST 2024 - MAMI				
	F1					F1				
Model	Macro	Micro	WA	ICM Norm	Mean	Macro	Micro	WA	ICM Norm	Mean
Baseline: RoBERTa+ Swin Transformer V2	0.43	0.50	0.51	0.29	0.43	**0.50**	**0.56**	**0.63**	0.37	**0.52**
SVM Style-Emo	0.46	0.52	0.52	0.31	0.45	0.47	0.55	0.61	**0.38**	0.50
RoBERTa	**0.47**	**0.53**	0.52	0.31	**0.46**	0.48	0.55	0.62	**0.38**	0.51
Ensemble	0.46	**0.53**	**0.53**	**0.32**	**0.46**	0.48	**0.56**	**0.63**	**0.38**	0.51

Table 4.2: In-domain and cross-dataset results for the multimodal baseline, component models, and ensemble in multi-label classification. WA: Weighted-averaged F1 score.

component models. For misogyny identification, RoBERTa outperformed the other models, including the ensemble; however, both performed equally well in detecting misogynous memes.

In the cross-dataset setup, the ensemble strategy clearly outperformed all component models in sexism identification. However, in misogyny identification, it only outperformed the best individual model in terms of macro and micro F1 scores. While the confusion matrices confirmed that the ensemble made the fewest errors in cross-dataset misogyny identification, it failed to correctly identify more misogynous memes than the multimodal baseline model.

4.1.2 Multi-label Classification

Table 4.2 presents the results of multi-label classification across the two evaluation settings explored, in-domain and cross-dataset. The results are described below according to each setting. The confusion matrices for multi-label classification that will be introduced below (Figures 4.5, 4.6, 4.8 and 4.7) present the average predictions across all classes, including the negative class at the binary level and the fine-grained classes. The class-specific confusion matrices are provided in Appendix B.

In-Domain

The results for multi-label classification when training and testing on each dataset individually showed that while the ensemble resulted in a good performance across both datasets, it did not outperform the component models.

Starting with the EXIST 2024 dataset, the fine-tuned RoBERTa model was the best component model, with a mean score of 0.52 and the best macro F1 score of 0.50. However, SVM model achieved the best weighted-average F1 of 0.64 across the individual models, outperforming the deep learning models in this metric and demonstrating

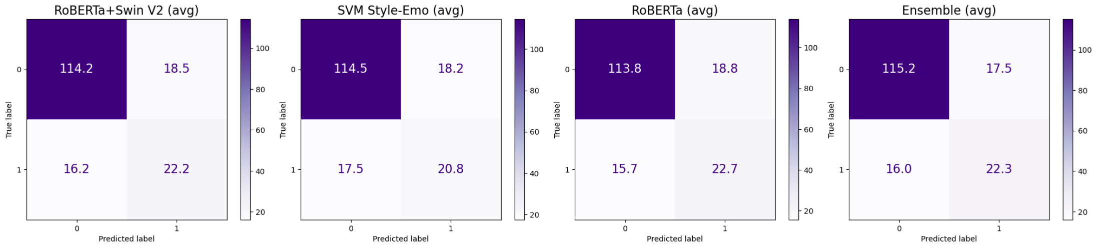


Figure 4.5: Confusion matrices for the component models and ensemble approach in multi-label classification on EXIST 2024. avg: Average of predicted fine-grained classes.

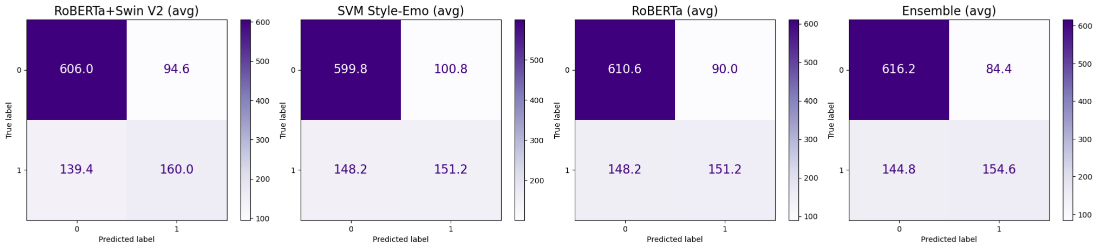


Figure 4.6: Confusion matrices for the component models and ensemble approach in multi-label classification on MAMI. avg: Average of predicted fine-grained classes.

the strength of the text-only conventional machine learning approach for sexism categorization. While ensemble strategy also yielded a mean score of 0.52 as RoBERTa, it resulted in the best ICM score of 0.37. This indicates that it was the most effective in categorizing sexism hierarchically.

Figure 4.5 presents the averaged confusion matrices for in-domain sexism categorization. The ensemble strategy made the fewest errors, with an average of 33.5 (16 FN and 17.5 FP), followed by RoBERTa with an average of 34.5 errors (15.7 FN and 18.8 FP). This helps explain why the ensemble achieved a higher ICM metric, even though RoBERTa was better at predicting the positive instances of fine-grained classes. Despite being a minimal difference, the SVM model presented the smallest rate of FP, with 18.2 across all models, possibly contributing to the reduction of FP in the ensemble in the in-domain setup, as occurred in cross-dataset sexism identification.

On the MAMI dataset, the multimodal baseline combining Swin Transformer V2 and RoBERTa resulted in the best overall performance across all evaluation metrics among the component models, with a mean score of 0.56. The ensemble strategy achieved the same scores as the baseline for most of the metrics, except for micro F1, for which the multimodal model outperformed it with 0.58 vs. 0.57. Figure 4.6 presents the averaged confusion matrices for in-domain misogyny categorization. While the ensemble was the model that exhibited the fewest mistakes, with a total average of 229.2 (144.8 FN and 84.4 FP), the multimodal baseline model predicted the most positive instances of fine-grained misogynous classes.

The results obtained on the MAMI dataset were higher than those on EXIST, which is reasonable given that the former contained fewer fine-grained categories (four, compared to five in EXIST 2024). In addition, the MAMI dataset counted with more training instances than EXIST 2024, which had 78% fewer training instances than MAMI in the fine-grained classes. Nonetheless, the models still achieved a decent performance in multi-label classification on both datasets, considering the complexity of the task, as was highlighted by the IAA achieved on the MAMI dataset for fine-grained classes. Regarding the confusion matrices, it can be observed that all models were mostly biased towards the negative class across both datasets, missing the true fine-grained instances. This can be attributed to the fact that they were exposed to more negative than positive instances per label during training, as the the number of memes in each-fine grained class was lower than the number of negative memes.

Cross-Dataset

The cross-dataset experiments yielded varying performance across the datasets, but the ensemble did not outperform the component models in none of them when considering the mean score.

When training on MAMI and testing on EXIST 2024, RoBERTa achieved the best mean score of 0.46 for sexism categorization among the component models. It also produced the highest macro and micro F1 scores, at 0.50 and 0.57 respectively. The ensemble strategy matched RoBERTa in mean and micro F1 scores. While it did not surpass RoBERTa in macro F1, it achieved the highest ICM Norm score of 0.32 and weighted F1 of 0.53. Notably, both RoBERTa and the ensemble strategy also reached the highest mean scores in the in-domain setting, with RoBERTa attaining the best macro F1, and the ensemble the best ICM Norm.

Regarding the performance difference between in-domain and cross-dataset sexism categorization, RoBERTa fine-tuned for sexism categorization showed a decline of 0.06 points in the mean score, with other metrics dropping between 0.03 to 0.09. The ensemble experienced the same drop in mean score, as both models performed identically on that metric. For the remaining metrics, its performance declined was consistent with the range from RoBERTa. The SVM model exhibited the largest drop in weighed F1, decreasing by 0.12 points, but the smallest in the mean score, with a drop of 0.04. Interestingly, the performance of the conventional machine learning classifier did not decline in macro F1. This highlights its strength in the categorization of sexism in multimodal memes, in particular in a cross-dataset setting.

Figure 4.7 shows the confusion matrices for cross-dataset sexism categorization. While the ensemble made an average of 45.3 errors (18.5 FN and 26.8 FP), RoBERTa correctly captured more positive instances of fine-grained classes. Nonetheless, SVM produced the fewest mistakes, with an average of 45 misclassification (19.5 FN and 25.5 FP). It yielded the lowest number of FP across the component models again, reinforcing the claim that they contribute to a reduced FP rate in the ensemble (Markov and Daelemans, 2021), in this case, for cross-dataset sexism categorization in memes.

When the EXIST 2024 dataset was used for training and the MAMI dataset for testing, the multimodal baseline achieved the highest mean score of 0.52 among the component models. It outperformed the other individual models in most of the metrics, except for ICM Norm, in which SVM and RoBERTa attained a score of 0.38. While the ensemble did not outperform the multimodal baseline, it did result in identical micro and weighted F1, with 0.56 and 0.63 respectively, and an ICM Norm of 0.38.

The drop in performance on this dataset was smaller than on the EXIST 2024, with the multimodal baseline showing a decline of 0.04 points and the ensemble 0.05 points in the mean score. Across this score, the SVM had the smallest drop in performance,

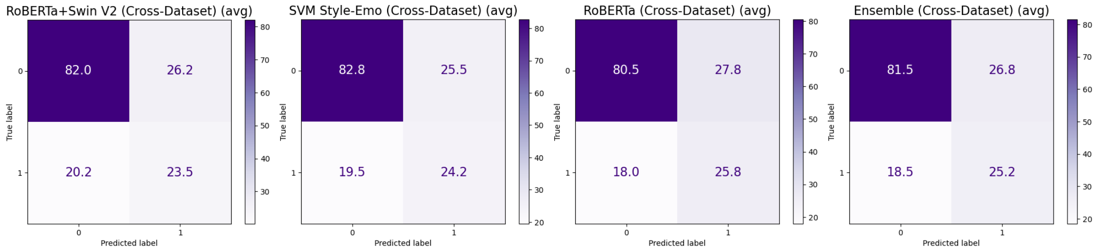


Figure 4.7: Confusion matrices for the component models and ensemble approach in multi-label classification when training on MAMI and testing on EXIST 2024. avg: Average of predicted fine-grained classes.

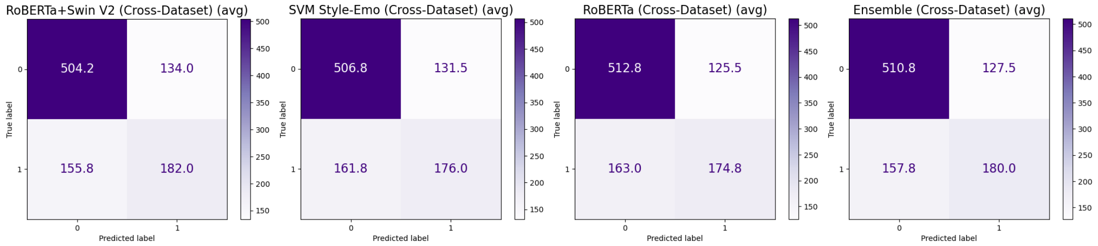


Figure 4.8: Confusion matrices for the component models and ensemble approach in multi-label classification when training on EXIST 2024 and testing on MAMI. avg: Average of predicted fine-grained classes.

with a decrease of 0.02 points. This text-only model also showed no drop in the micro F1 score, highlighting its robustness as the only system that maintained performance, at least on this metric, reinforcing its strength in a cross-dataset setup, as previously demonstrated (Markov et al., 2021; Markov and Daelemans, 2021). Figure 4.8 presents the confusion matrices for cross-dataset misogyny categorization. The ensemble was the model with the fewest errors, with an average of 285.3 (157.8 FN and 127.5 FP), but the multimodal baseline remains the system with the most correctly predicted memes for the fine-grained classes.

Summary

In the multi-label classification task, the ensemble performed well in both in-domain and cross-dataset setups. While it did not outperform the best component models across all metrics, it showed improvements in the ICM metric by making fewer errors. However, the component models were more effective at correctly predicting fine-grained instances. RoBERTa was the best-performing model for sexism categorization in both setups, while the multimodal baseline combining Swin Transformer V2 with RoBERTa achieved the best performance for in-domain and cross-dataset misogyny categorization. The SVM model also delivered consistent results across evaluation setups, demonstrating its capabilities for the categorization of sexism and misogyny in memes in in-domain and cross-dataset setups. In addition, it contributed to a lower FP rate in the ensemble in in-domain and cross-dataset sexism categorization, even though the ensemble was not the best-performing approach overall.

Unlike in binary classification, the results for multi-label classification were not consistent across datasets, despite their similar sizes. As observed in the in-domain setup, the MAMI dataset exhibited the smallest drop in performance in the cross-dataset setting. A possible explanation for the larger performance drop on the EXIST dataset in this setup could be its small test set size, which offers little margin for error.

4.2 Qualitative Analysis

This section analyses the predictions of ensemble strategies as they are the main focus of this thesis. A correlation analysis will provide insight into the contributions of the component models, while a detailed error analysis will help explain how ensemble strategies are helpful in the detection of sexism and misogyny in memes, addressing their limitations and thereby the research questions.

4.2.1 Correlation Analysis

The correlation of the predictions was measured using the Pearson correlation coefficient. This metric assesses the linear relationship between two variables, in this case the predictions of each model. It outputs a value between -1 and +1, where 0 indicates no correlation, and -1 and +1 indicate a perfect negative or positive linear relationship, respectively. A positive correlation implies that the two variables move in the same direction, while a negative correlation means that they move in different directions. Correlations where r is close to 0 are considered weaker than those closer to +1 or -1. The formula is provided below:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y)}}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where x_i and y_i refer to the data points of variables X and Y, and $\bar{x}$ and $\bar{y}$ represent their respective means. The formula computes the covariance of X and Y, divided by the product of their standard deviations. The result indicates the strength and direction of their linear relationship. The Pearson correlation coefficient was implemented with the *stats*[1] library from SciPy (Virtanen et al., 2020).

As Table 4.3 presents, the predictions of the component models exhibited a moderate positive degree of correlation across all evaluation settings. For the binary classification task in the in-domain setting, the SVM produced the least correlated predictions for sexism identification on the EXIST 2024 dataset. In contrast, the deep learning models showed a higher degree of correlation, with an r value of 0.68. On the MAMI dataset, the predictions from the SVM and the multimodal model were the least correlated for misogyny identification, compared to the stronger correlation observed between the deep learning models. In the cross-dataset setup, the predictions from the SVM model were again the most uncorrelated across both datasets.

The correlation values for multi-label classification followed the same pattern as those for binary classification. In the experiments on the EXIST 2024 dataset, the predictions from the SVM model were the least correlated for sexism categorization. This was most evident in the in-domain setting when compared to the predictions from RoBERTa, and in the cross-dataset setting when compared to those of the multimodal

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html

		In-Domain		Cross-Dataset	
	Predictions	EXIST 2024	MAMI	MAMI - EXIST 2024	EXIST 2024 - MAMI
Binary Classification	SVM vs. RoBERTa	**0.40**	0.52	**0.39**	**0.36**
	SVM vs. RoBERTa+Swin Transformer V2	**0.30**	**0.44**	**0.27**	**0.34**
	RoBERTa vs. RoBERTa +Swin Transformer V2	0.68	0.58	0.61	0.54
Multi-label Classification	SVM vs. RoBERTa	**0.49**	0.52	0.56	**0.47**
	SVM vs. RoBERTa+Swin Transformer V2	0.53	**0.47**	**0.49**	**0.48**
	RoBERTa vs. RoBERTa+Swin Transformer V2	0.59	0.64	0.66	0.54

Table 4.3: Correlations between predictions of component models with Pearson correlation coefficient. The values in multi-label classification are averaged across classes, and those with $r < 0.5$ are highlighted in bold type.

model. A similar trend was observed on the MAMI dataset, where the SVM produced the least correlated predictions for misogyny categorization across both evaluation settings.

While RoBERTa and the multimodal baseline both have a deep learning architecture, they appeared to have learned similar patters, as indicated by their higher correlated predictions. For an ensemble to be more effective than its individual component models, they need to not only perform well, but also introduce diversity in their predictions (Sagi and Rokach, 2018) given that it was demonstrated that there is a linear relationship between error reduction rate and the degree of uncorrelated predictions among component models (Ali and Pazzani, 1995). Therefore, the less correlated predictions from the SVM introduced diversity to the ensemble, possibly cancelling out the errors of the other models. This, in turn, helped reduce the error rate of the ensemble and contributed to stronger overall performance, in particular on the EXIST 2024 dataset, where the ensemble outperformed all component models for binary sexism identification across the evaluation metrics. This contribution from SVM goes in line with the findings from previous research (Markov et al., 2021). However, it would probably have been more beneficial to the ensemble, especially in the multi-label classification task, if the predictions from at least two models were uncorrelated (or the least correlated, as in binary classification) since it was also proven that including the predictions from two uncorrelated models produces the greatest error reduction (Kyriakides and Margaritis, 2019).

4.2.2 Error Analysis

A detailed error analysis was performed to gain deeper insights into the performance of the ensemble strategies across the evaluation settings. Since the ensemble was the best-performing system in binary classification, the error analysis can help understand why they are effective for the identification of sexism and misogyny in memes. The analysis was inspired by van Aken et al. (2018), who identified and described common errors made by ensembles in the detection of harmful textual content. Their work focused on analysing false negative (FN) and false positive (FP) predictions to address gaps in precision and recall.

The errors of the ensembles were analysed for each evaluation setting, i.e. in-domain vs. cross-dataset, in the binary classification task only. Analysing the errors in the multi-label classification task would require a detailed examination of each category. Since this study focused on the average performance and the ensemble did not outperform the component models in this task, multi-label classification errors were not explored–also due to time constrains. Thirty memes were randomly selected per task and manually analysed, considering the meme itself, the meme text, and the image caption. In cases where the number of errors did not reach this figure, all available memes were examined. The most representative errors are described below.

Error Classes of False Negatives

The error classes of FN in binary classification were investigated for sexism and misogyny identification in both in-domain and cross-dataset setups. The most common type of errors found across the 99 memes manually inspected were: sexism or misogyny without swear or offensive words, rhetorical questions, metaphors and comparisons, sarcasm and irony, incorrect or poor suggestive image description, and the quality of the meme text that was provided with the dataset. While some memes might present more than one type of error, each error was assigned to a single category. Therefore, the statistics provided reflect individual errors only and do not account for intersections.

Sexism or misogyny without swear or offensive words This type of error refers to harmful memes that do not contain explicit hate or swear words. It was found found in both in in-domain and cross-dataset setups in sexism identification, but only in cross-dataset misogyny identification. This occurred in 16% and 15% of the sexism identification errors in in-domain and cross-dataset, and in 13% of the errors analysed for misogyny identification in the cross-dataset setting. The examples found in this category did not have a woman in the meme image, and only feminine pronouns in the meme text were found in some instances, which could be the reason why they were predicted as negative. This category had the second-highest occurrence in in-domain sexism identification, following *sarcasm and irony.*

Figure 4.9(a) presents an example of an error in this category, meme ID: 16305.jpg, which belongs to the MAMI test set. The corresponding image caption was: a man in a suit and tie. While there are not explicit swear words in the meme, the combination of image and the text made it a misogynistic instance that the ensemble missed. In this

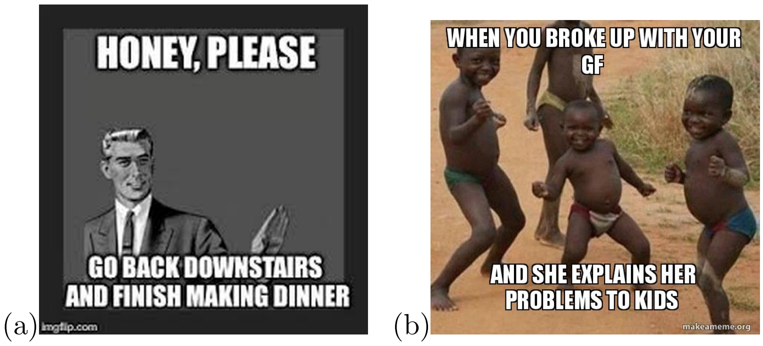

(a)
HONEY, PLEASE
GO BACK DOWNSTAIRS AND FINISH MAKING DINNER
imgflip.com
(b)
WHEN YOU BROKE UP WITH YOUR GF
AND SHE EXPLAINS HER PROBLEMS TO KIDS
makeameme.org

Figure 4.9: Examples of FN errors of the type sexism or misogyny without swear or offensive words.

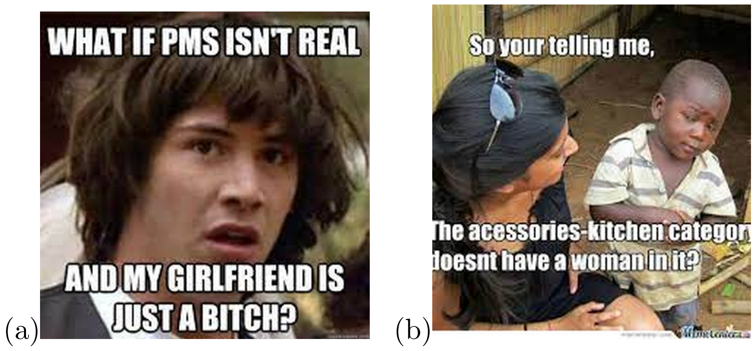

(a)
WHAT IF PMS ISN'T REAL
AND MY GIRLFRIEND IS JUST A BITCH?
(b)
So your telling me,
The acessories-kitchen categor
doesnt have a woman in it?

Figure 4.10: Examples of FN errors of the type rhetorical questions.

particular example, there is also an ironical tint in the text despite the use of polite words. In another example from the EXIST dataset (b), meme id: 210630, which was an false negative in both in-domain and cross-dataset, the image caption said: a group of children playing in the dirt. While the image caption missed the fact that the children are laughing, the meme would imply that the kids are laughing at the woman referred to in the meme text as “GF”, which stands for girlfriend.

Rhetorical questions As highlighted by van Aken et al. (2018), harmful content can be expressed through rhetoric or suggestive questions, or in text containing question words or question marks. This type of error was present in at least 11% of the errors in in-domain sexism identification, and 5% in the cross-dataset setting. It occurred in 23% of the sampled memes for in-domain misogyny identification, and in 3% in cross-dataset setup.

Figure 4.10 contains two examples of this error type. The examples are from the MAMI dataset, meme IDs 15348.jpg and 15159.jpg. Meme (a) was missed by the ensemble even though there was an offensive word in the meme text. Moreover, this particular meme contains an abbreviation, which could have made it more difficult to classify. Its image caption was “a man with a tie”, which misses the actual expression in the face of the person in the image. While meme (b), which image caption was “a woman and a child talking to each other”, did not have a question word as meme (a), it does have a question mark. In this category, some memes included women in the image, as meme (b), or in the text, as meme (a).

Metaphors and comparisons Processing metaphors and comparisons requires an understanding of implied meaning or additional world knowledge. This type of errors were found in the identification of sexism in in-domain and cross-dataset settings, with 11% and 10% of the FN in each of them. In misogyny identification, it occurred in 23% of the errors in in-domain and 10% in cross-dataset.

Figure 4.11 presents two examples from the MAMI test set (meme IDs 15187.jpg and 15232.jpg). In both cases, the comparison happens between the image and the text, a pattern also observed in the other memes found in this category. The first meme (a) implies that the character is attractive, while the second (b) suggests the person looks tired. Their respective image captions were “captain marvel in a red and blue outfit” and “a woman with a long hair and a hat”. Although the latter was not accurate in terms of the woman wearing a hat, a model would need to assess the “attractiveness”

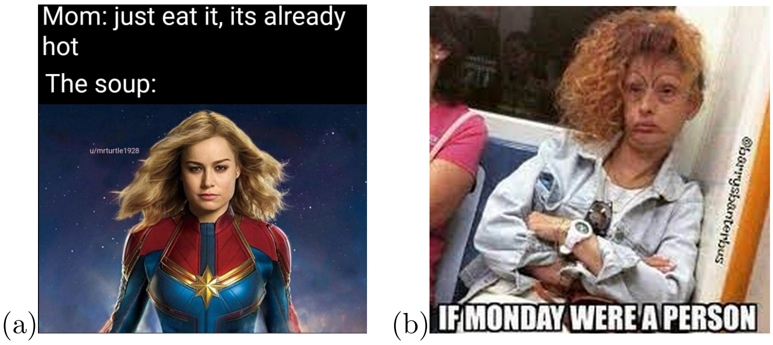


Figure 4.11: Examples of FN errors of the type metaphors and comparisons.

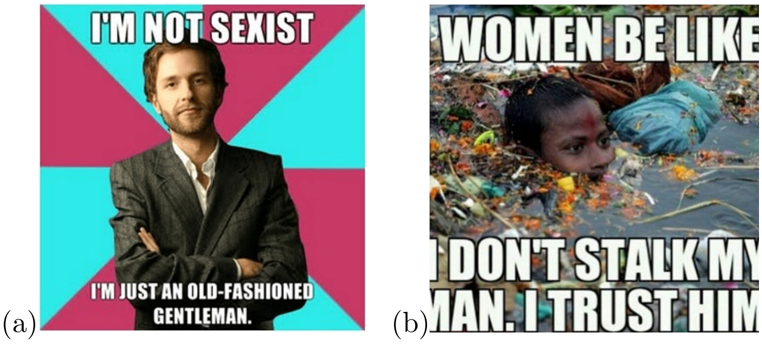

(a)
I'M NOT SEXIST
I'M JUST AN OLD-FASHIONED GENTLEMAN.
(b)
WOMEN BE LIKE
I DON'T STALK MY MAN. I TRUST HIM

Figure 4.12: Examples of FN errors of the type sarcasm and irony.

of a person in (a), or interpret or describe their features in a derogatory way in (b) to classify them correctly as misogynous, both of which are highly subjective tasks. Most memes in this category relied on intentionally mean comparisons that would require a very detailed image processing for a model to associate them with misogyny or sexism.

Sarcasm and irony Whenever sarcasm and irony are used as rhetorical devices, they can obscure the intended meaning (van Aken et al., 2018) since the meme usually means the opposite of what is stated in the meme text. This makes the correct identification of sexism and misogyny in memes more challenging. This FN error occurred in 42% and 45% of sexism identification errors in in-domain and cross-dataset settings, respectively. It was also present in 30% of in-domain misogyny identification errors and in 30% of cross-dataset errors. This was the category with the most FN across both datasets and evaluation setups, highlighting the need to address sarcasm and irony in the identification of sexist and misogynous memes. In many cases, the sarcasm or irony was more evident when the image conveyed a meaning opposite to that of the text.

Figure 4.12 presents two examples of this type of error, with meme IDs 211375 (a) and 210909 (b) from the EXIST 2024 dataset. The image caption in (a) was "a man with his arms crossed". This meme was a false negative in both in-domain and cross-dataset setups, confirming that irony was hard to grasp even when training with a different dataset. Moreover, upon examining the image caption in (b), which was "a boy in the water", it failed to adequately describe the "predator-like" position of the person in the water, as if observing prey. This highlights the lack of detailed descriptions in the image captions obtained with BLIP-2.

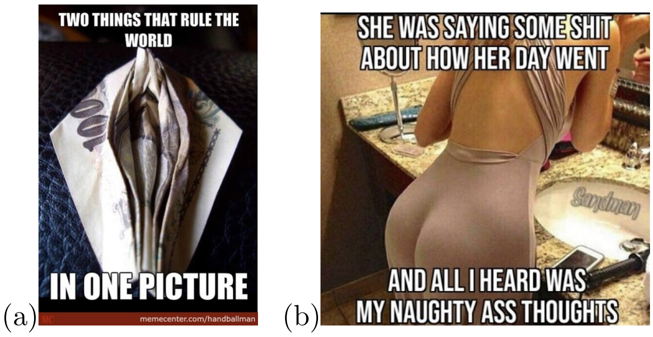

(a)
TWO THINGS THAT RULE THE WORLD
IN ONE PICTURE
memecenter.com/handballman
(b)
SHE WAS SAYING SOME SHIT ABOUT HOW HER DAY WENT
AND ALL I HEARD WAS MY NAUGHTY ASS THOUGHTS

Figure 4.13: Examples of FN errors of the type suggestive or explicit image.

Incorrect or poor suggestive image description Upon inspecting the errors, a novel class emerged as many cases did not fit into the previously defined error classes but shared a common feature: a suggestive or explicit image. In these instances, while many memes were combined with forms of irony, the image itself was the primary source of the sexism or misogyny, leaving the irony or sarcasm as a secondary trait. It was also notable that the image captions that were obtained with BLIP-2 were often uninformative. This type of error mostly occurred in misogyny identification, accounting for 30% of in-domain errors and 36% of cross-dataset errors, being the second highest error in in-domain misogyny identification after *sarcasm and irony*, and the highest in the cross-dataset setup. In contrast, it was observed in 11% of the errors in in-domain sexism identification and only 5% in cross-dataset. Since it mostly occurred in MAMI, this could be due to the larger size of the dataset, whereas such explicit memes might not have been present in the EXIST 2024 dataset due to its inherently smaller size. This could also explain why the errors in cross-dataset misogyny identification were higher than in the in-domain setting in the samples analysed.

Given the wrong predictions and the subpar quality of the image captions, it is also assumed that the visual embeddings that were processed by the multimodal model were not able to capture the level of detail from such explicit and suggestive images. Examples of this error from the MAMI dataset are provided in Figure 4.13. Meme (a) (ID 15541.jpg) was found in in-domain and (b) (ID 15073.jpg) in cross-dataset errors. In (a), the image caption was "a folded piece of money", while the banknote is folded in a way that resembles a vagina. The caption in (b) was "a woman in a dress". To address this type of error, both the model used to produce image captions and the model used to generate visual embeddings in the multimodal approach would need to be highly descriptive in a rude or unpleasant manner in order to correctly classify these instances as misogynous.

Quality of the meme text It was also noted that the quality of the meme text provided in the datasets was subpar in some instances. This refers not to the text on the meme itself, but to the textual data included as part of the dataset for processing. This error often appeared in combination with other type of errors, such as rhetorical question. Moreover, some memes grouped under this category contained more than one image. In sexism identification, this error occurred in 10% and 20% of the FN in in-domain and cross-dataset settings, respectively. In misogyny, in contrast, it was only found in 3% of the errors in both settings. This was the second most-common error in cross-dataset sexism identification after *sarcasm and irony*.

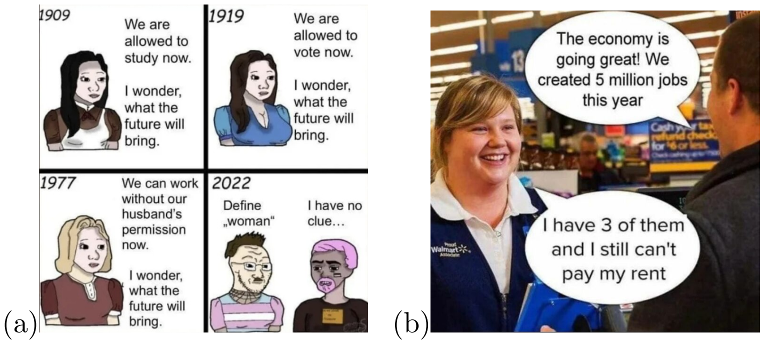

(a)
1909
We are allowed to study now.
I wonder, what the future will bring.
1919
We are allowed to vote now.
I wonder, what the future will bring.
1977
We can work without our husband's permission now.
I wonder, what the future will bring.
2022
Define „woman"
I have no clue...
(b)
The economy is going great! We created 5 million jobs this year
I have 3 of them and I still can't pay my rent

Figure 4.14: Examples of FN errors of the type quality of the meme text.

Figure 4.14 presents two memes from the EXIST 2024 dataset. The meme text for meme (a) (meme ID 211875) was "1909 1977 We are allowed to study now. I wonder, what the future will bring. We can work without our husband's permission now. I wonder, what the future will bring. 1919 2022 Define "woman" We are allowed to vote now. I wonder, what the future will bring. I have no clue...". In this meme, the text did not follow the order of each box from left to right and from top to bottom, and instead appeared in an unordered way from left to right. In (b), (meme ID 211031) the meme text also included text from the background posters and the logo in the uniform of the lady, as in: Proud Walmart. Astocate The economy is going great! We created 5 million jobs this year Cash y tax check for 6 or less 10 I have 3 of them and I still can't pay my rent ins. Moreover, the order in which the text appeared follows a left to right order. As shown in these examples, instances with wrong meme text were difficult to label during evaluation, and they could have introduced noise, especially if they were also present in the training dataset.

Error Classes of False Positives

The false positive predictions were carefully examined across in-domain and cross-dataset settings for sexism and misogyny identification. A total of 114 memes were analysed, and the following categories were identified: doubtful labels, use of swear or offensive words, women in meme, men in meme, domestic chores, and reference to violence.

Doubtful labels It was observed that some memes labelled as non-misogynous in the MAMI dataset might, in fact, be considered misogynous according to the definition provided by the dataset authors. This occurred in 3% of the examined memes in the in-domain setting and 16% in the cross-dataset setting. While the dataset included examples of memes with transphobic content labelled as misogynous, some similar memes were labelled as non-misogynous. The FP in cross-dataset suggests that the transphobic memes in the EXIST 2024 dataset were more consistently labelled, as the ensemble was able to learn and correctly identify them as positive instances of misogyny in this case. This category was not found among the FP on EXIST 2024 dataset.

Figure 4.15 presents two examples of this error type. (a) (meme ID 15772.jpg) is referring to feminism as a movement that spread lies and conceptually depicts hateful content in the form of stereotype. In (b) (meme ID 15725.jpg), in contrast, the text appears to denigrate the person in the meme. It is possible that meme (b) refers to a spe-

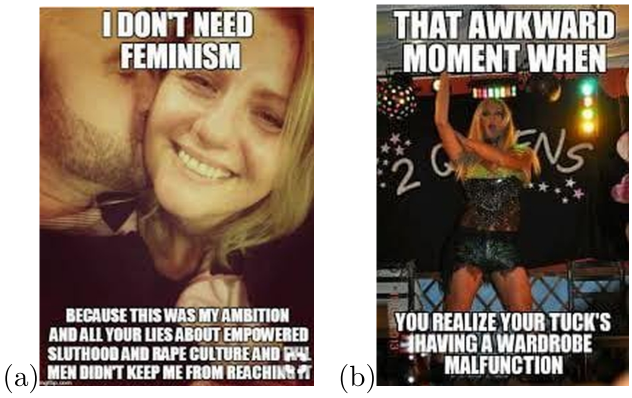

(a)
I DON'T NEED FEMINISM
BECAUSE THIS WAS MY AMBITION AND ALL YOUR LIES ABOUT EMPOWERED SLUTHOOD AND RAPE CULTURE AND EVIL MEN DIDN'T KEEP ME FROM REACHING IT
imgflip.com
(b)
THAT AWKWARD MOMENT WHEN
YOU REALIZE YOUR TUCK'S HAVING A WARDROBE MALFUNCTION

Figure 4.15: Examples of FP errors of the type doubtful labels.

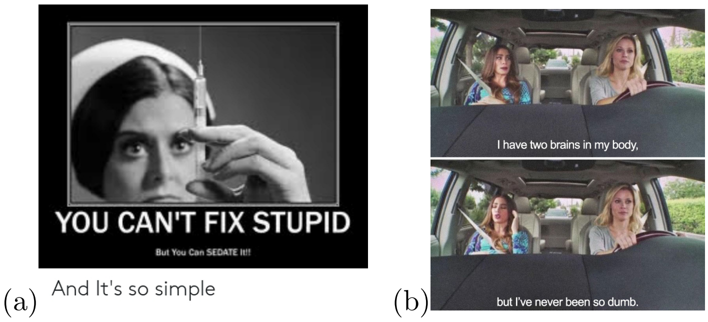


Figure 4.16: Examples of FP errors of the type swear or offensive words.

cific character, which would require world knowledge to recognize it is a non-misogynous instance; therefore, it remains a doubtful case. These two examples highlight potential label misalignment in the MAMI dataset. Given its large size (11k memes), some cases of incorrect labelling could have occurred.

Use of swear or offensive words As highlighted by van Aken et al. (2018), the ensemble might have learned that the use of swear or offensive words are a sign of misogyny and sexism. However, a problem raises when memes use these words in a non-misogynous or non-sexist way. In most memes under this category, offensive words either appeared with pictures of women or included words typically used against them, such as “bitch” or “whore”. This error occurred in 16% of FP in in-domain sexism identification, and 13% in cross-dataset. As for misogyny identification, this type of error was found in 10% of in-domain FP and 3% in cross-dataset.

Figure 4.16 illustrates this error type with two examples: (a) from the MAMI dataset (meme ID 16304.jpg), and (b) from the EXIST 2024 dataset (meme ID 211336). Their respective image captions were “a nurse holding a syringe” and “two women in a car”. These examples suggest that the component models in the ensemble learned to associate the use of swear words and images of women with positive instances of sexism or misogyny.

Women in meme Apart from the errors involving offensive language in the memes, some memes containing non-sexist or non-misogynous text were incorrectly flagged as positive instances. This mainly occurred in memes featuring close-up shots of women or

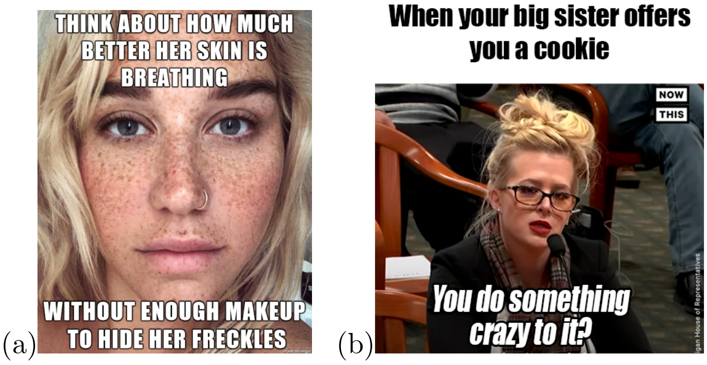

THINK ABOUT HOW MUCH BETTER HER SKIN IS BREATHING
WITHOUT ENOUGH MAKEUP TO HIDE HER FRECKLES
(a)
When your big sister offers you a cookie
NOW THIS
You do something crazy to it?
(b)

Figure 4.17: Examples of FP errors of the type women in meme.

girls, mostly alone but sometimes in groups. Moreover, this also occurred with memes in which women were mentioned in the meme text, as in “your mother”, or “other women”. This error was found in 42% of the FP in in-domain sexism identification, and in 47% of the cross-dataset setup. Most of the memes that were misclassified in the cross-dataset setup were also misclassified in in-domain. The error also occurred in misogyny identification, with 47% and 37% in each setting. Among all these memes, 15% depicted blonde women, 15% weight and body appearances, and 5% referred to makeup. This was the most common type of FP, probably because sexism and misogyny (often) target women. As a result, the ensemble was exposed to many examples of women in memes in which, for example, there were references to their appearance.

Two examples of this FP error type are shown in Figure 4.17. Meme (a) belongs to the MAMI dataset, with meme ID 15049.jpg, and (b) to the EXIST 2024 dataset, with meme ID 211484. The image caption of (a) was “a girl with freckles”, and of (b) “a woman with glasses”. Both of them were misclassified in in-domain and cross-dataset experiments. This type of error suggests that the visual modality was stronger than the textual modality.

Men in meme Just as some FP featured memes only about women, others included only an image of a man accompanied by non-sexist or non-misogynous text, but in most cases, the text still referred to women. Men images were also mostly close-ups, as occurred in the errors related to women in meme. In sexism identification, this error accounted for 17% of FP and 13% of cross-dataset errors, while in misogyny identification, it appeared in 20% of in-domain errors and 10% of cross-dataset errors. This could suggest that the EXIST 2024 dataset was more suitable for the ensemble to learn to correctly classify these instances compared to the MAMI dataset, since there were less errors of this type found in the cross-dataset setup. Interestingly, 25% of the memes in this FP class referred to “cheating”.

Figure 4.18 presents two examples of this error with (a) from EXIST 2024 (meme ID 210901) and (b) from MAMI (meme ID 16233.jpg). Their respective image captions were “a man with a beard and glasses” for (a), and “a man with a ring on his finger” for (b). While the latter was not very accurate, it still mentioned a man. This type of error might emerge from training memes in which men were shown with a sexist or misogynous caption in which women were mentioned.

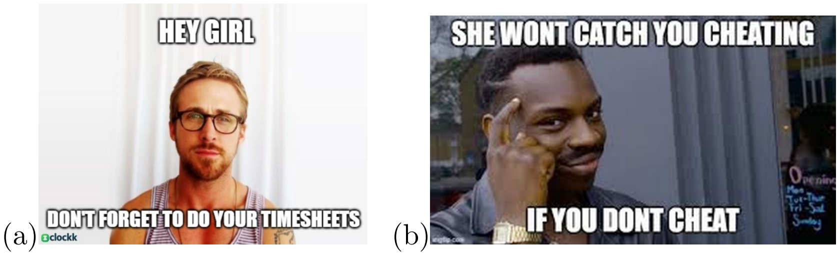

(a)
HEY GIRL
DON'T FORGET TO DO YOUR TIMESHEETS
clockk
(b)
SHE WONT CATCH YOU CHEATING
IF YOU DONT CHEAT

Figure 4.18: Examples of FP errors of the type men in meme.

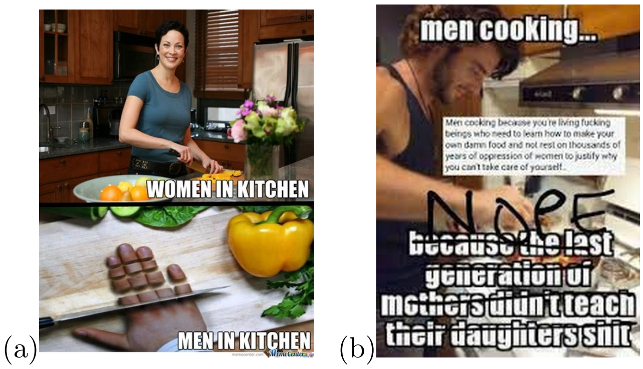

(a)
WOMEN IN KITCHEN
MEN IN KITCHEN
(b)
men cooking...
Men cooking because you're living fucking beings who need to learn how to make your own damn food and not rest on thousands of years of oppression of women to justify why you can't take care of yourself.
NOPE
because the last generation of mothers didn't teach their daughters shit

Figure 4.19: Examples of FP errors of the type domestic chores.

Domestic chores Another type of error involved memes containing references to domestic chores. This FP error occurred mostly in misogyny identification, with 10% in in-domain and 13% in cross-dataset. In sexism identification, it only happened in 6% of cross-dataset misclassifications. The memes contained images or mentions of cooking appliances or kitchens, with both women and men involved. The system might have learned to associate references to domestic chores, both in text and images, with positive instances of sexism or misogyny due to similar examples in the training data.

Figure 4.19 presents two examples of this error type. Memes (a) and (b) from MAMI the dataset (ID 15137.jpg and 15358.jpg) contain references to men in the kitchen. In (a) a woman is compared to a man (image caption: a woman in a kitchen with a knife and a man), and in (b) a man is cooking (image caption: a man cooking on a stove). These errors occurred in the cross-dataset setup, meaning that they were predicted as positive instances when the EXIST 2024 dataset was used for training. In example (a), sexism is directed at men through irony in the combination of image and text. In (b), however, the meme conveys the opposite message but was still predicted as a positive instance of misogyny. Apart from these cases, there were memes that contained mentions of women in the kitchen but then contradicted that stereotype. It appears that the ensemble focused on the domestic chore reference to label the memes as sexist or misogynous, failing to correctly classy them as negative instances.

Reference to violence The final category identified for FP errors includes memes containing words or images that suggest violent actions or scenarios, with words such as “kill” or “terrorist”, or images like black eyes or explosions, combined with depictions of women or girls, even though the memes did not actually contain sexism or misogyny. This error occurred in 17% of in-domain and 13% of cross-dataset FP in sexism identification. In misogyny identification, it was found in 6% of in-domain errors and

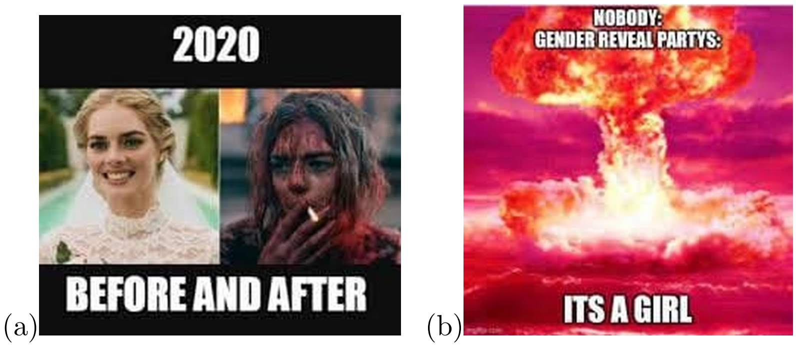


Figure 4.20: Examples of FP errors of the type reference to violence.

16% of cross-dataset. This error might originate from all the instances in the training dataset in which women were displayed in violent scenes either in the picture or in the text.

Figure 4.20 presents two examples from this FP error class, both from MAMI dataset, with meme ID 15977.jpg and 15309.jpg. While the image caption in (a) was not very informative (a woman with a cigarette and a woman with a cigarette), the model was able to capture the meaning of the image of a woman covered in blood as hateful, suggesting the strength of the visual embeddings. In (b), however, the image caption was accurate (a nuclear explosion with a pink background), but the combination of the explosion and the word "girl" appears to have led the ensemble to predict it as a misogynous meme.

Summary of Errors

After conducting a detailed error analysis of the patterns found in the false negative and false positive predictions, it was determined that the use of rhetorical devices such as sarcasm and irony in memes resulted in the highest percentage of false negative across both datasets in in-domain classification. While this category was also the most common cause of false negatives in cross-dataset sexism identification, a suggestive image in the meme was the cause of the most false negatives in cross-dataset misogyny identification. This might originate from a poor description of the image both in image captioning and visual embeddings from the multimodal model, which is something that could be addressed by future work. Other error categories were sexism or misogyny without swear or offensive words, rhetorical questions, metaphors and comparisons, and the quality of the meme text.

Regarding false positives, the presence of women in memes, either in the image or the text, was the most common trigger for misclassification of sexism and misogyny, both in in-domain and cross-dataset setups. The other categories identified in FP errors for both sexism and misogyny were the use of swear or offensive words, images of men in the meme, and a reference to violence. Some instances of doubtful labels were also found in the MAMI dataset, and a reference to domestic chores was identified in in-domain and cross-dataset misogyny identification, as well as cross-dataset sexism identification.

Chapter 5

Conclusion and Discussion

5.1 Conclusions

This study focused on the identification and categorization of sexism and misogyny in memes, with evaluation in in-domain and cross-dataset setups. Since ensembles of conventional machine learning and deep learning models have proven useful for the classification of textual harmful content, the same approached was explored with the aim of determining whether such a method would also work for the identification and categorization of sexism and misogyny in multimodal content.

The methodology implemented to address this research consisted of a hard majority voting ensemble strategy that combined conventional machine learning and deep learning models. The datasets used for the experiments were the MAMI (Multimedia Automatic Misogyny Identification) from SemEval-2022 Task 5 and the training set from EXIST (sEXism Identification in Social neTworks), introduced at CLEF 2024. The tasks involved binary classification in which memes were classified as sexist or misogynous or not, depending on the dataset used, and a hierarchical multi-label classification task in which the fine-grained categories of sexism and misogyny were identified. Apart from evaluating the performance of each dataset in the typical in-domain setup, this study also implemented a cross-dataset evaluation to determine how well the models generalize to another task, resembling real-world applications. This required a stratified sampling technique to keep the overlapping classes for multi-label classification.

The models combined in the ensemble strategy included two models based on the textual modality, i.e. the meme text and image caption, and a state-of-the-art multimodal model. The text-only models were SVM with stylometric and emotion-based features, experimented with across various configurations, and a pre-trained language model. For the latter, different models were tested, with RoBERTa ultimately selected as the component model used throughout the experimental setups. Moreover, a multimodal model was implemented both as a component model in the ensemble and as a baseline to assess the performance improvements introduced by the ensemble. After exploring different language and visual models, the final architecture consisted of Swin Transformer V2 combined with RoBERTa, followed by an MLP fusion model and prediction layer for classification.

The results showed that an ensemble strategy of conventional machine learning and deep learning models yielded the best performance in the binary task of sexism identification, both in in-domain and cross-dataset setups. Although the ensemble matched the overall performance of the text-only RoBERTa model in multi-label sexism categorization, it was not the best-performing system in any of the evaluation settings. This also occurred in the binary and multi-label tasks of misogyny identification and categorization using the MAMI dataset, where the results with the ensemble were similar to those of the best component models in both in-domain and cross-dataset scenarios. The most common misclassifications patterns made by the ensemble in binary classification were identified through an error analysis. The use of rhetorical devices such as sarcasm and irony accounted for the highest percentage of false negatives across both datasets in in-domain classification. In contrast, incorrect or poor suggestive image descriptions were the primary source of false negatives in cross-dataset misogyny identification. Regarding false positives, the presence of women in memes, either in the image or in the text, was the most frequent cause across both datasets in in-domain and cross-dataset settings.

5.1.1 Addressing Research Questions

This thesis focused around the research question:

RQ: Given that ensemble strategies have shown promising results for detecting harmful content in textual data, including in cross-domain setups, are they also beneficial for the detection and classification of sexism and misogyny in multimodal data?

The following sub-questions were posed to answer the main research question:

1. What are the best component models to be incorporated into the ensemble?

The best component models to be incorporated into the ensemble were identified by exploring different models and configurations of text-only and multimodal models on the EXIST 2024 dataset, since it was smaller and faster to process. In terms of text-only models, an ablation study was performed on different configurations of an SVM model with stylometric and emotion-based features, and different pre-trained models were explored and fine-tuned, including BERT, RoBERTa, and a BERTweet fine-tuned for sexism detection. The model that resulted in the best performance for sexism identification on the test set was RoBERTa. In a similar way, different models were explored for the multimodal approach, with ViT and Swin Transformer V2 for extracting the visual features, and BERT and RoBERTa for extracting the textual features. The combinations explored were BERT with Swin Transformer V2, RoBERTa with Swin Transformer V2, and BERT with ViT. After testing these models on both development and test sets, the approach that combined RoBERTa with Swin Transformer V2 was the best-performing approach. Therefore, it was chosen as the multimodal baseline and third component model to apply across all experimental setups on both datasets.

2. Is the ensemble strategy helpful for both binary and multi-label classification across different datasets?

3. Is the ensemble strategy useful for the in-domain setup as well as cross-dataset setting?

The ensemble strategy, which combined conventional machine learning and deep learning models, proved helpful for binary classification by outperforming the individual component models in the identification of sexism, but not in the identification of misogyny. In the multi-label classification task, however, the ensemble did not achieve better performance than the component models for either sexism or misogyny categorization.

In the task where the ensemble strategy performed best–binary sexism identification–it was useful in both in-domain and cross-dataset setups by being the top-performing system. In the binary classification task of misogyny identification, as well as in the multi-label classification tasks of misogyny and sexism categorization, the ensemble did not outperform the component models in any of the evaluation settings, neither in-domain nor cross-dataset.

4. Given that shallow approaches like SVM help reduce the false positive rate in text, do they also contribute to a lower false positive rate in multimodal data when incorporated into an ensemble?

The SVM with stylometric and emotion-based features produced the lowest number of false positives among the component models in the binary task of sexism identification in the cross-dataset setup. Although the difference was minimal compared to the other individual models, the SVM also had the lowest false positive rate in both in-domain and cross-dataset multi-label sexism categorization, even though the ensemble was not the best-performing models in these settings. This behaviour was not observed in any of the experiments related to misogyny identification or categorization. Since the SVM clearly contributed to a lower false positive rate in only one out of four experiments for binary classification (in-domain sexism identification), further research using additional datasets would be required to verify whether it consistently helps to reduce false positives in an ensemble for the classification of memes.

5. Based on error and correlation analyses, why are ensemble strategies helpful?

A correlation analysis revealed that the predictions from the SVM were the least correlated compared with those of the deep learning models in the ensemble, across both binary and multi-label classification tasks, and in both in-domain and cross-dataset settings. In the binary setting, the correlations between the SVM with the pre-trained and multimodal models were weaker, indicating that they were less correlated, which introduced more diversity into the ensemble and helped to mitigate the errors of the other models. A detailed error analysis of the binary classification predictions in both in-domain and cross-dataset settings revealed the most common patters of misclassifications made by the ensemble. For false negatives, the presence of sarcasm and irony accounted for most errors across both datasets in the in-domain setting. In cross-dataset misogyny identification, the main source of false negatives was incorrect or poor suggestive image descriptions. As for false positives, the most common cause across both datasets, in both in-domain and cross-dataset evaluation setups, was the presence of women in memes, either in the image or in the text.

According to the answer to each sub-question, it can be deduced that ensemble strategies are beneficial for the binary classification of sexism in memes. While this was not the case for misogyny identification in memes, an ensemble of conventional machine learning and deep learning models outperformed the component models in in-domain and cross-dataset binary sexism identification.

5.2 Limitations

The scope and outcomes of this research were influenced by a number of limiting factors. Starting with the datasets, while the size of the EXIST 2024 dataset allowed for quick processing during the initial experiments, its small size proved limiting, particularly because the test labels were not available for experimentation and the results could not be compared to the state-of-the-art. Furthermore, by starting the experiments with this datasets, precisely because of its ease of processing, the configuration and selection of the models was focused on it. This was because the size of the MAMI dataset did not allow for experimentation of models across both datasets to determine which would work best for both, given that running the experiments on this dataset required longer processing time and GPUs on Google Colaboratory platform. This could be a reason for the lower performance in the in-domain setting in binary classification, for example, in particular for the multimodal approach.

The experimental setup selected for multi-label classification was based on binary relevance, which might have had an impact on the performance since it does not consider a relationship between labels. Moreover, the class imbalance of each fine-grained class against the negative class might have an impact of the performance in multi-label classification.

The selection of models was limited to open source pre-trained models, rather than closed-source LLMs, due to time and resource constraints. As a result, the implemented methods focused on open-source pre-trained language models, both for preprocessing (image captioning) and as component models in the ensemble. In terms of component models, the use of RoBERTa and the Swin Transformer V2 combined with RoBERTa resulted in higher prediction correlations, likely due to their similar encoder-based architectures. Considering that ensembles tend to perform better when at least two produce uncorrelated predictions, having only one model with uncorrelated predictions might limit the overall performance gains that the ensemble could achieve.

Another limitation arose from the selection of model used to generate the image captions. Although different prompts were explored on the open source vision-language model BLIP-2, the error analysis revealed that the resulting image captions were often uninformative and, in many cases, inaccurate. Since the implemented SVM model relied on stylometric and emotion-based features, short or uninformative captions might have failed to provide meaningful input for these features. The error analysis also showed that suggestive or explicit images were the main cause of false positives in misogyny classification. While the image captions lacked sufficient detail regarding the sexual content of the images, the results also suggested that the visual embeddings generated by the multimodal model were also not sufficiently informative for detecting this type of content.

Finally, the quality of the meme text in the memes in the EXIST 2024 dataset was another limitation. Combined with inaccurate or poorly descriptive image captions, this negatively affected model performance, particularly because all models relied on the meme text provided in the dataset. Moreover, doubtful labels in the MAMI dataset posed a further challenge, as they might have led the models to learn incorrect patterns during training, based on the high recall of the misogynous class.

5.3 Future Work

The limitations identified above open up several opportunities for future research in the identification and categorization of sexism and misogyny in memes. Starting with the datasets, the MAMI dataset could be revised for doubtful labels to make sure that the models learn the correct patterns from the training data. Moreover, the quality of the meme text in the EXIST dataset could also be checked, in particular in memes that are composed by multiple images.

In terms of model selection, the optimal feature configuration for the stylometric and emotion-based SVM could be explored on the MAMI dataset. Given that this research focused on optimizing the features using the EXIST 2024 dataset to ensure generalizability across all evaluation scenarios, this approach might have limited the performance of the SVM model on the MAMI dataset.

Moreover, replacing the SVM classifier with another transformer-based model could be beneficial to address the tasks at hand. Even though the SVM produced uncorrelated predictions, including an additional encoder-only model could improve the overall performance of the ensemble. In addition, the use of LLMs as component models could be explored. An ensemble tends to perform better when it includes at least two models with uncorrelated predictions; therefore, adding a model that performs well on the task while offering uncorrelated outputs could further boost performance. Exploring different architectures than the ones implemented in this research, such as decoder-based models, might bring further improvements to the classification of sexist and misogynous memes.

The image captions could be generated with another vision-language model. Using a model that can provide a detailed description of the images would be beneficial, in particular, one that can accurately describe more explicit or suggestive images, as it was confirmed that they posed the most errors in misogyny identification. In the same line, the features resulting from the vision model in the multimodal approach could be explored to confirm whether they are indeed able to capture this level of detail in suggestive images. The challenge presented by explicit images could also be approached by implementing NSFW (not safe/suitable for work) features that capture indecent content, such as pornographic images or drawings, sexually explicit images such as bikini or nude photos, and safe for work neutral images and drawings[1] (Rehman et al., 2025). This might help differentiate content in which women are displayed sexually from those where they are not, possible helping to reduce both false negatives and false positives in an ensemble.

Since rhetoric devices, such as sarcasm, irony, metaphors, and comparisons, proved to be recurrent sources of error in multimodal sexism and misogyny identification, future work could address them as auxiliary tasks within one of the component models of the ensemble. It would be crucial to approach these subtasks in a multimodal manner, as the errors in this research showed that the meaning of these rhetoric devices often emerges only through the relationship between the text and the image.

Last but not least, the number of negative instances used during training for multi-label classification could be balanced to account for the class imbalance between the fine-grained classes and the negative class, as the models showed a bias toward the latter. In addition, multi-label classification could be explored using an alternative approach in which each combination of labels is treated as a unique label. An error

[1] https://github.com/alex000kim/nsfw_data_scraper

analysis of this task could be conducted in future research focusing on the false positives and false negatives to uncover the causes of misclassification across each fine-grained class, in both in-domain and cross-dataset settings, as was done for binary classification in this research.

Appendix A

Preprocessing: Image Captioning

A.1 Phrases Removed in Image Captions

Phrase	**MAMI**	**EXIST 2024**	**Both**
with the words	51	7	58
with the caption	336	75	411
with a caption	1,952	371	2,323
with the text	28	5	33
with a text	35	10	45
with a quote	92	23	115
with a message	60	7	67
with text saying	35	4	39
with text that says	37	6	43
and a caption	561	129	690
and a sign that says	10	0	10
and a quote that says	1	0	1
and a text that says	1	0	1
and a text saying	7	2	9
and text that says	5	1	6
and the caption says	1	0	1
and saying	7	1	8
with a funny meme	1	1	2
Total	3,220	642	3,862

Table A.1: Total phrases removed from image captions across datasets.

Appendix B

Multi-Label Classification: Confusion Matrices

B.1 In-Domain

B.1.1 EXIST 2024

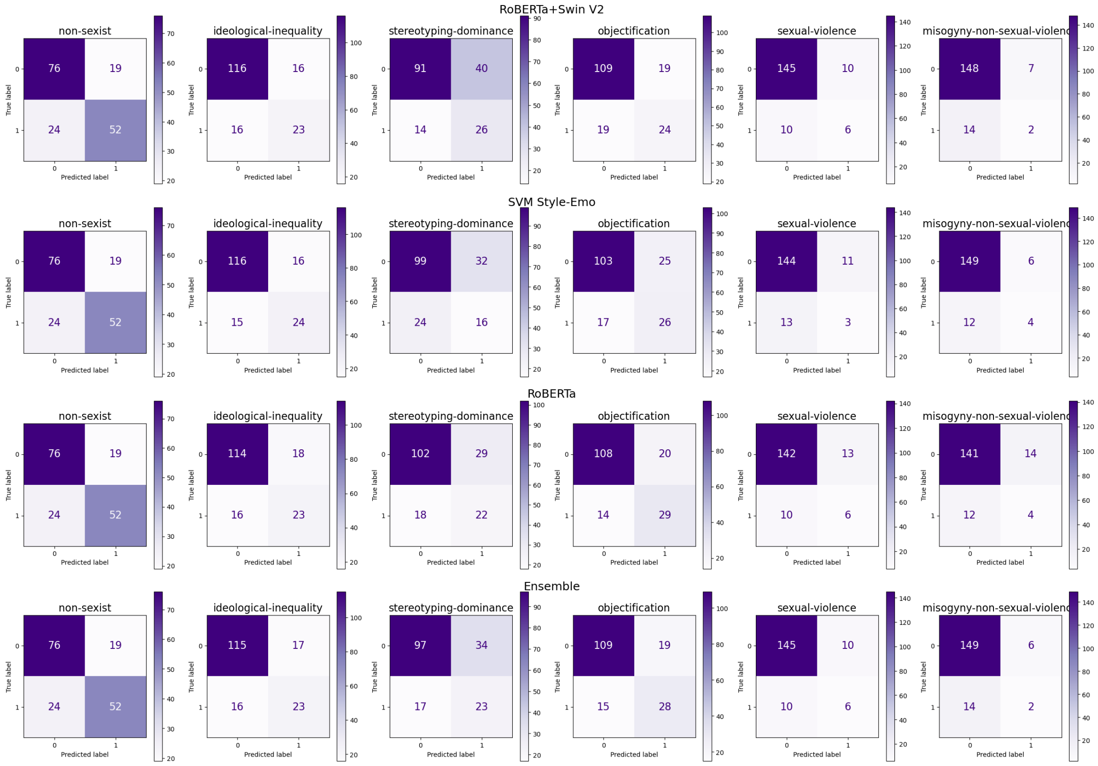


Figure B.1: Confusion matrices per fine-grained category per model implemented on the EXIST 2024 dataset. Non-sexist category predicted with Ensemble in the first step.

B.1.2 MAMI

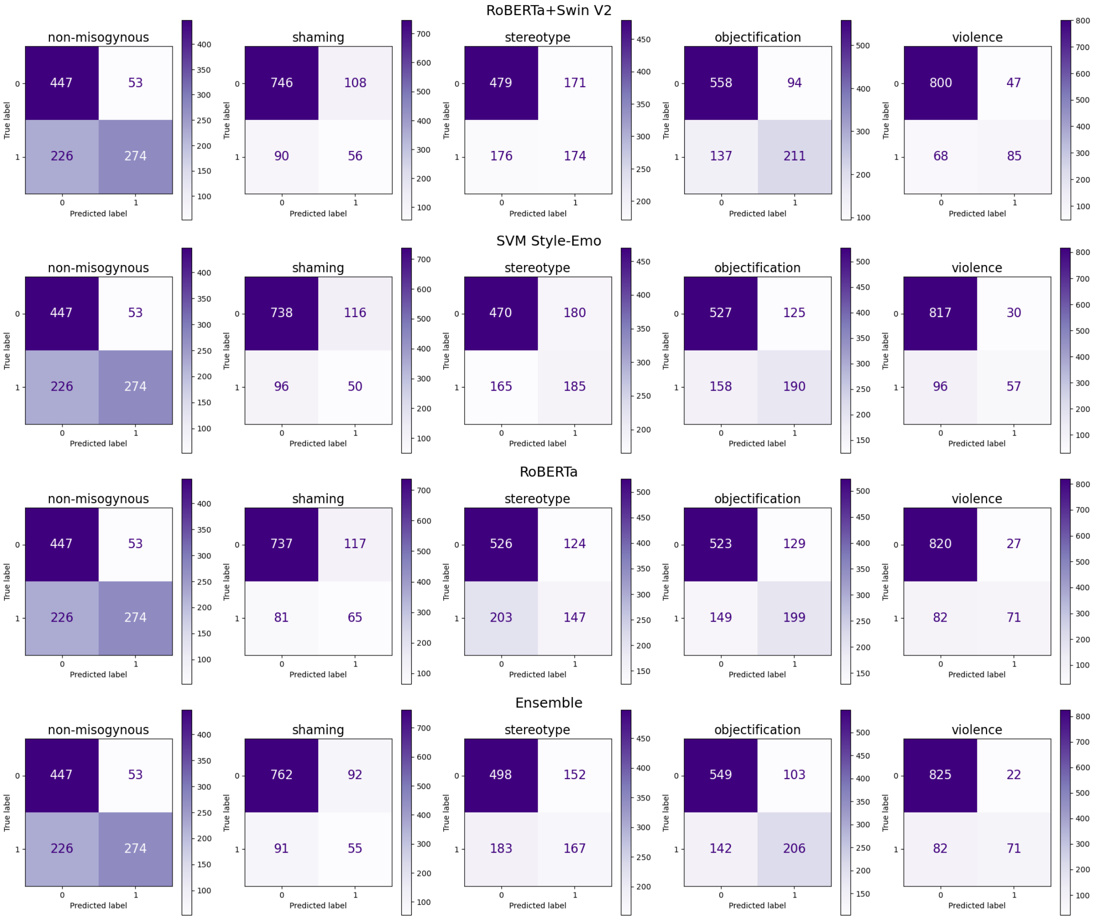


Figure B.2: Confusion matrices per fine-grained category per model implemented on the MAMI dataset. Non-misogynous category predicted with Ensemble in the first step.

B.2 Cross-Dataset

B.2.1 EXIST 2024

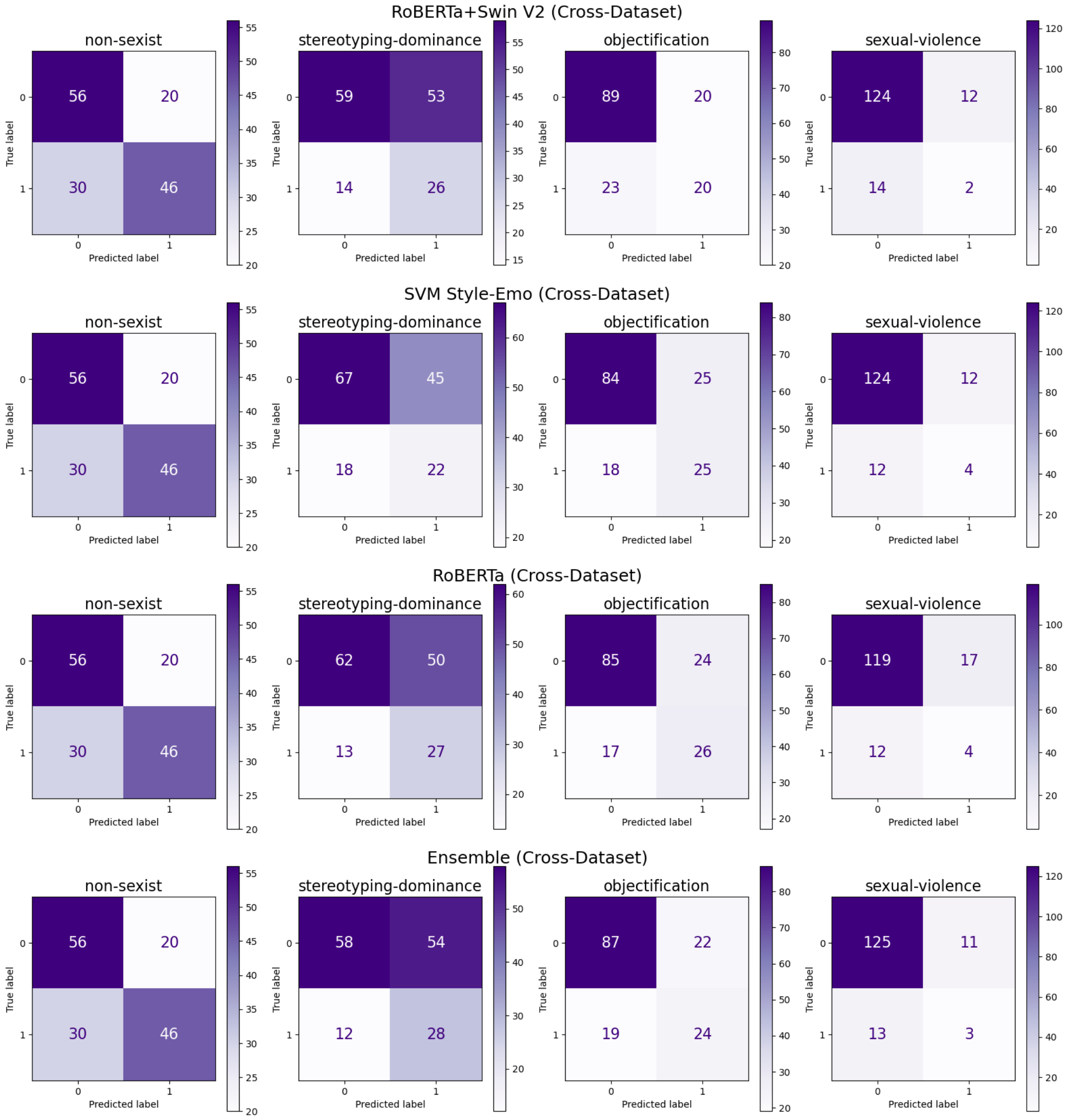


Figure B.3: Confusion matrices per fine-grained category per model when training on MAMI dataset and testing on EXIST 2024 dataset. Non-sexist category predicted with Ensemble in the first step.

B.2.2 MAMI

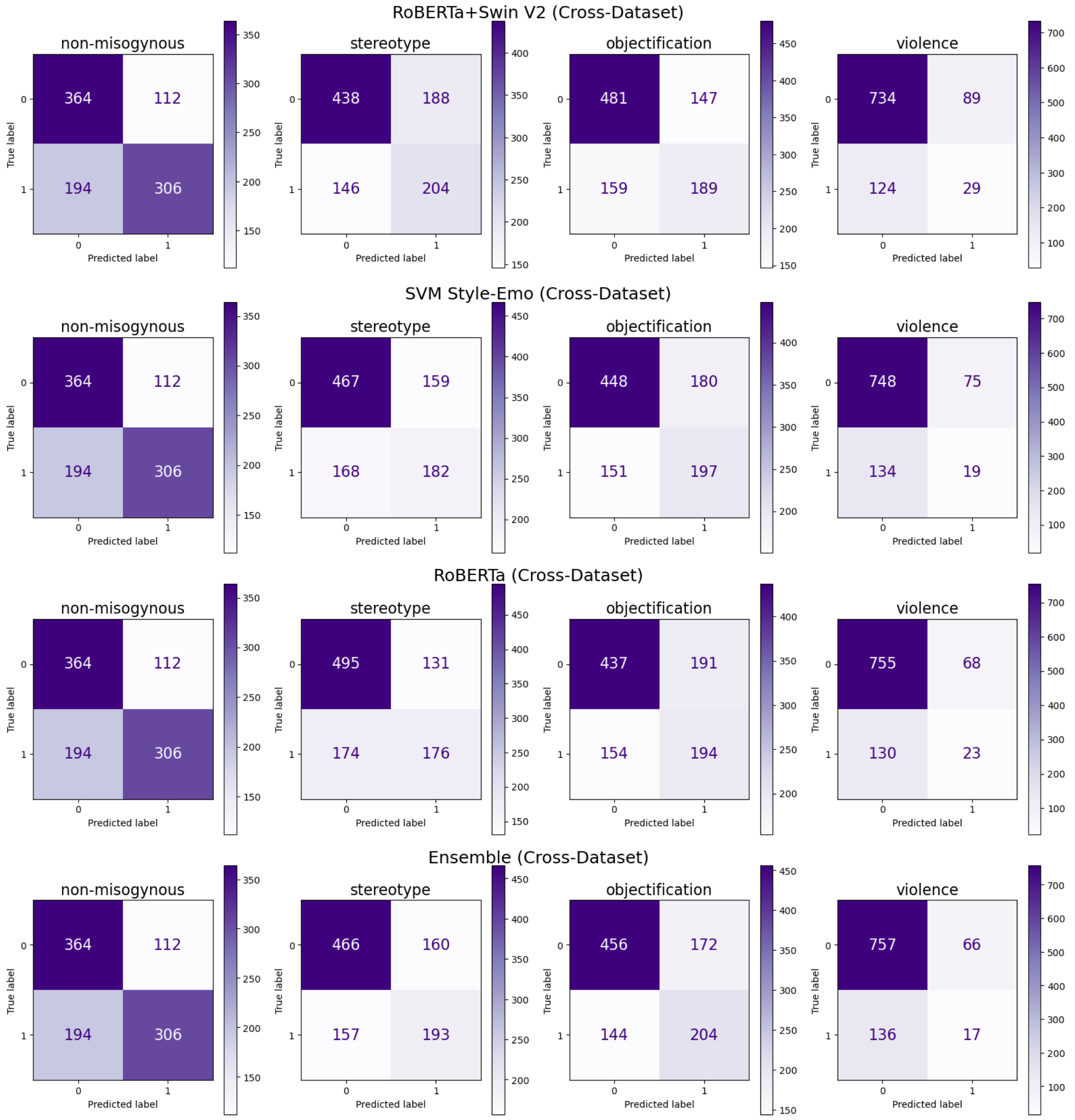


Figure B.4: Confusion matrices per fine-grained category per model when training on EXIST 2024 dataset and testing on MAMI dataset. Non-misogynous category predicted with Ensemble in the first step.

Bibliography

P. Aggarwal, J. Mehrabanian, W. Huang, Ö. Alacam, and T. Zesch. Text or Image? What is More Important in Cross-Domain Generalization Capabilities of Hate Meme Detection Models? In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 104–117, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.8/.

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330701. URL https://doi.org/10.1145/3292500.3330701.

S. Al-Azzawi, G. Kovács, F. Nilsson, T. Adewumi, and M. Liwicki. NLP-LTU at SemEval-2023 task 10: The Impact of Data Augmentation and Semi-Supervised Learning Techniques on Text Classification Performance on an Imbalanced Dataset. In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1421–1427, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.196. URL https://aclanthology.org/2023.semeval-1.196/.

K. M. Ali and M. J. Pazzani. On the Link between Error Correlation and Error Reduction in Decision Tree Ensembles. Technical report, UC Irvine: Donald Bren School of Information and Computer Sciences, 1995. URL https://escholarship.org/uc/item/9k55v622.

E. Amigó and A. Delgado. Evaluating Extreme Hierarchical Multi-label Classification. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.399. URL https://aclanthology.org/2022.acl-long.399/.

M. Anzovino, E. Fersini, and P. Rosso. Automatic identification and classification of misogynistic language on twitter. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, and F. Meziane, editors, *Natural Language Processing and Information Systems*, pages 57–64, Cham, 2018. Springer International Publishing. ISBN 978-3-319-91947-8.

E. Bassignana, V. Basile, V. Patti, et al. Hurtlex: A Multilingual Lexicon of Words to Hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253, pages 1–6. CEUR-WS, 2018.

S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

D. Cameron. *Language, Sexism and Misogyny*. Routledge, London, 1 edition, Nov. 2023. ISBN 978-1-003-29411-5. doi: 10.4324/9781003294115. URL https://www.taylorfrancis.com/books/9781003294115.

J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*, 2020.

C. Cervone, M. Augoustinos, and A. Maass. The Language of Derogation and Hate: Functions, Consequences, and Reappropriation. *Journal of Language and Social Psychology*, 40(1):80–101, 2021. doi: 10.1177/0261927X20967394.

Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80, 2012. doi: 10.1109/SocialCom-PASSAT.2012.55.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

M.-C. de Marneffe, C. D. Manning, J. Nivre, and D. Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07 2021. ISSN 0891-2017. doi: 10.1162/coli_a_00402.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

M. Duggan. Online Harassment 2017, July 2017. URL https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2017/07/PI_2017.07.11_Online-Harassment_FINAL.pdf.

R. Faris, A. Ashar, U. Gasser, and D. Joo. Understanding Harmful Speech Online. Networked Policy Series 2016-18, Berkman Klein Center for Internet & Society, Dec. 2016. URL `https://cyber.harvard.edu/publications/2016/UnderstandingHarmfulSpeech`.

E. Fersini, D. Nozza, and P. Rosso. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *EVALITA Evaluation of NLP and Speech Tools for Italian Proceedings of the Final Workshop 12-13 December 2018, Naples*. Accademia University Press, Dec. 2018a.

E. Fersini, P. Rosso, and M. Anzovino. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *IberEval@ sepln*, 2150:214–228, Sept. 2018b.

E. Fersini, F. Gasparini, and S. Corchs. Detecting Sexist MEME On The Web: A Study on Textual and Visual Cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231, 2019. doi: 10.1109/ACIIW.2019.8925199.

E. Fersini, D. Nozza, P. Rosso, et al. AMI@ EVALITA2020: Automatic Misogyny Identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. CEUR-WS, 2020.

E. Fersini, G. Rizzi, A. Saibene, and F. Gasparini. Misogynous MEME Recognition: A Preliminary Study. In *AIxIA 2021 – Advances in Artificial Intelligence: 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 1–3, 2021, Revised Selected Papers*, page 279–293, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-031-08420-1. doi: 10.1007/978-3-031-08421-8_19.

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/`.

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378, 1971. doi: 10.1037/h0031619.

S. Frenda, B. Ghanem, M. Montes-y Gómez, and P. Rosso. Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752, May 2019. ISSN 10641246, 18758967. doi: 10.3233/JIFS-179023.

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/`.

M. Hasanain, M. A. Hasan, F. Ahmad, R. Suwaileh, M. R. Biswas, W. Zaghouani, and F. Alam. ArAIEval shared task: Propagandistic techniques detection in unimodal and multimodal Arabic content. In N. Habash, H. Bouamor, R. Eskander, N. Tomeh, I. Abu Farha, A. Abdelali, S. Touileb, I. Hamed, Y. Onaizan, B. Alhafni, W. Antoun, S. Khalifa, H. Haddad, I. Zitouni, B. AlKhamissi, R. Almatham, and K. Mrini, editors, *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 456–466, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.arabicnlp-1.44. URL https://aclanthology.org/2024.arabicnlp-1.44/.

P. He, X. Liu, J. Gao, and W. Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention, 2021. URL https://arxiv.org/abs/2006.03654.

P. He, J. Gao, and W. Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2023. URL https://arxiv.org/abs/2111.09543.

H. Jarquín-Vásquez, I. Tlelo-Coyotecatl, M. Casavantes, D. I. Hernández-Farías, H. J. Escalante, L. Villaseñor-Pineda, M. Montes, et al. Overview of DIMEMEX at IberLEF 2024: Detection of Inappropriate Memes from Mexico. *Procesamiento del Lenguaje Natural*, 73:335–345, 2024.

D. Jurafsky and J. H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 2025. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released January 12, 2025.

U. Khurana, I. Vermeulen, E. Nalisnick, M. Van Noorloos, and A. Fokkens. Hate Speech Criteria: A Modular Approach to Task-Specific Hate Speech Definitions. In K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, and Z. Talat, editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.17. URL https://aclanthology.org/2022.woah-1.17/.

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. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33 of *NIPS '20*, pages 2611–2624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.

H. Kirk, W. Yin, B. Vidgen, and P. Röttger. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.305. URL https://aclanthology.org/2023.semeval-1.305/.

G. Kyriakides and K. G. Margaritis. *Hands-on Ensemble Learning with Python: build highly optimized ensemble machine learning models using scikit-learn and Keras.*

Packt Publishing, Birmingham, UK, 2019. ISBN 978-1-78961-285-1, 978-1-78961-788-7.

A. Lees, J. Sorensen, and I. Kivlichan. Jigsaw@AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT model. In V. Basile, D. Croce, M. Di Maro, and L. C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, 2020.

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.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al. Swin Transformer v2: Scaling Up Capacity and Resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022. doi: 10.1109/CVPR52688.2022.01170.

J. Ma and R. Li. RoJiNG-CL at EXIST 2024: Sexism Identification in Memes by Integrating Prompting and Fine-tuning. In G. Faggioli, N. Ferro, P. Galuščáková, and A. García Seco de Herrera, editors, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, 2024.

I. Markov and W. Daelemans. Improving Cross-Domain Hate Speech Detection by Reducing the False Positive Rate. In A. Feldman, G. Da San Martino, C. Leberknight, and P. Nakov, editors, *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4if-1.3. URL https://aclanthology.org/2021.nlp4if-1.3/.

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/.

I. Markov, I. Gevers, and W. Daelemans. An Ensemble Approach for Dutch Cross-Domain Hate Speech Detection. In *International Conference on Applications of*

Natural Language to Information Systems, pages 3–15, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-08472-0. doi: 10.1007/978-3-031-08473-7_1.

Y. Mehdad and J. Tetreault. Do Characters Abuse More Than Words? In R. Fernandez, W. Minker, G. Carenini, R. Higashinaka, R. Artstein, and A. Gainer, editors, *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, Los Angeles, Sept. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3638. URL https://aclanthology.org/W16-3638/.

A. Menárguez Box and D. Torres Bertomeu. DiTana-PV at sEXism Identification in Social neTworks (EXIST) Tasks 4 and 6: The Effect of Translation in Sexism Identification. In G. Faggioli, N. Ferro, P. Galuščáková, and A. García Seco de Herrera, editors, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, 2024.

S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013. doi: 10.1111/j.1467-8640.2012.00460.x.

D. Q. Nguyen, T. Vu, and A. Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.2. URL https://aclanthology.org/2020.emnlp-demos.2/.

M. Paciello, F. D'Errico, G. Saleri, and E. Lamponi. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in Human Behavior*, 116:106655, 2021. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2020.106655.

R. Passonneau. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation.* European Language Resources Association (ELRA), 01 2006.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, and V. Ruiz. EXIST 2024: sEXism Identification in Social neTworks and Memes. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, and I. Ounis, editors, *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V*, page 498–504, Berlin, Heidelberg, 2024a. Springer-Verlag. ISBN 978-3-031-56068-2. doi: 10.1007/978-3-031-56069-9_68.

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

(Extended Overview). In L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, page 93–117, 2024b. ISBN 978-3-031-71907-3. doi: 10.1007/978-3-031-71908-0_5.

B. Poland. *Haters: Harassment, abuse, and violence online.* Potomac Books, an imprint of the University of Nebraska Press, 2016. ISBN 9781612348728.

S. Ramis, J. M. Buades, F. J. Perales, and C. Manresa-Yee. A Novel Approach to Cross dataset studies in Facial Expression Recognition. *Multimedia Tools and Applications*, 81(27):39507–39544, Nov 2022. ISSN 1573-7721. doi: 10.1007/s11042-022-13117-2.

M. Z. U. Rehman, S. Zahoor, A. Manzoor, M. Maqbool, and N. Kumar. A Context-aware Attention and Graph Neural Network-based Multimodal Framework for Misogyny Detection. *Information Processing I& Management*, 62(1):103895, 2025. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2024.103895. URL https://www.sciencedirect.com/science/article/pii/S0306457324002541.

F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, and L. Plaza. Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access*, 8:219563–219576, 2020. doi: 10.1109/ACCESS.2020.3042604.

F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67:195–207, 2021.

O. Sagi and L. Rokach. Ensemble Learning: A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018. doi: 10.1002/widm.1249.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

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/.

K. Sechidis, G. Tsoumakas, and I. Vlahavas. On the stratification of multi-label data. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-23808-6.

X. Shi, J. Mueller, N. Erickson, M. Li, and A. Smola. Multimodal AutoML on Structured Tables with Text Fields. In *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021. URL https://openreview.net/forum?id=OHAIVOOl7Vl.

P. Szymański and T. Kajdanowicz. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*, Feb. 2017.

S. Thapa, K. Rauniyar, F. Jafri, H. Veeramani, R. Jain, S. Jain, F. Vargas, A. Hürriyetoğlu, and U. Naseem. Extended Multimodal Hate Speech Event Detection During Russia-Ukraine Crisis - Shared Task at CASE 2024. In A. Hürriyetoğlu, H. Tanev, S. Thapa, and G. Uludoğan, editors, *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 221–228, St. Julians, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.case-1.31/.

Y. Theocharis, S. Kosmidis, J. Zilinsky, F. Quint, and F. Pradel. Content Warning: Public Attitudes on Content Moderation and Freedom of Expression. *Content Moderation Lab at TUM Think Tank*, 2025. doi: 10.17605/OSF.IO/F56BH.

B. van Aken, J. Risch, R. Krestel, and A. Löser. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, editors, *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5105. URL http://aclweb.org/anthology/W18-5105.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Y. Wang and I. Markov. CLTL@Multimodal Hate Speech Event Detection 2024: The Winning Approach to Detecting Multimodal Hate Speech and Its Targets. In A. Hürriyetoğlu, H. Tanev, S. Thapa, and G. Uludoğan, editors, *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 73–78, St. Julians, Malta, Mar. 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.case-1.9/.

Y. Wang and I. Markov. CLTL at ArAIEval Shared Task: Multimodal Propagandistic Memes Classification Using Transformer Models. In N. Habash, H. Bouamor, R. Eskander, N. Tomeh, I. Abu Farha, A. Abdelali, S. Touileb, I. Hamed, Y. Onaizan, B. Alhafni, W. Antoun, S. Khalifa, H. Haddad, I. Zitouni, B. AlKhamissi, R. Almatham, and K. Mrini, editors, *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 501–506, Bangkok, Thailand, Aug. 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.arabicnlp-1.51. URL https://aclanthology.org/2024.arabicnlp-1.51/.

Y. Wang and I. Markov. CLTL at DIMEMEX Shared Task: Fine-Grained Detection of Hate Speech in Memes. In S. M. Jiménez-Zafra, L. Chiruzzo, F. Rangel, F. Balouchzahi, U. B. Corrêa, A. Bonet Jover, H. Gómez-Adorno, J. Á. González Barba, D. I. Hernández Farías, A. Montejo Ráez, P. Moral, C. Rodríguez Abellán, M. E. Vallecillo Rodríguez, M. Taulé, and R. Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*, 2024c.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2010. URL https://aclanthology.org/S19-2010/.

M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. SemEval-2020 task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.188. URL https://aclanthology.org/2020.semeval-1.188/.

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/.

M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12:191–202, 2018. doi: 10.1007/s11704-017-7031-7.

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/.

Z.-H. Zhou. *Machine Learning*. Springer Singapore, Singapore, 2021. ISBN 978-981-15-1966-6, 978-981-15-1967-3. doi: 10.1007/978-981-15-1967-3. URL https://link.springer.com/10.1007/978-981-15-1967-3.

R. Zhu. Enhance Multimodal Transformer with External Label and In-Domain Pretrain: Hateful Meme Challenge Winning Solution, 2020. URL https://arxiv.org/abs/2012.08290.