UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

INSTITUTO DE COMPUTAÇÃO- ICOMP

PROGRAMA PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

Enhancing Aggressive Content Detection in Memes Using Multimodal Machine Learning Models

Paulo Cezar de Queiroz Hermida

Manaus

January 2025

UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

INSTITUTO DE COMPUTAÇÃO- ICOMP

PROGRAMA PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

Paulo Cezar de Queiroz Hermida

Enhancing Harmful Content Detection in Memes Using Multimodal Machine Learning Models

Thesis presented to the Graduate Program in Informatics of the Institute of Computing of the Federal University of Amazonas in partial fulfillment of the requirements for the degree of Doctor in Informatics.

Advisor:

Eulanda Miranda dos Santos, Ph.D.

Manaus

January 2025

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Queiroz Hermida, Paulo Cezar de
Q3e Enhancing harmful content detection in memes using multimodal machine learning models / Paulo Cezar de Queiroz Hermida . 2025
128 f.: il. color; 31 cm.

Orientadora: Eulanda Miranda dos Santos
Tese (Doutorado em Informática) - Universidade Federal do Amazonas.

1. Multimodal machine learning. 2. Hate speech detection. 3. Generative models. 4. Content moderation. I. Santos, Eulanda Miranda dos. II. Universidade Federal do Amazonas III. Título

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"ENHANCING AGGRESSIVE CONTENT DETECTION IN MEMES USING MULTIMODAL MACHINE LEARNING MODELS"

PAULO CEZAR DE QUEIROZ HERMIDA

Tese de Doutorado defendida e aprovada pela banca examinadora constituída pelos professores:

Profa. Dra. Eulanda Miranda dos Santos - **Presidente**

Prof. Dr. André Luiz da Costa Carvalho - **Membro Interno**

Prof. Dr. Eduardo James Pereira Souto - **Membro Interno**

Prof. Dr. Marco Antônio Pinheiro de Cristo - **Membro Externo**

Prof. Dr. Raoni Simões Ferreira - **Membro Externo**

Manaus, 31 de janeiro de 2025.

Documento assinado eletronicamente por **Eulanda Miranda dos Santos**, **Professor do Magistério Superior**, em 06/02/2025, às 17:45, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Raoni Simões Ferreira**, **Usuário Externo**, em 11/02/2025, às 09:39, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Eduardo James Pereira Souto**, **Professor do Magistério Superior**, em 11/02/2025, às 13:40, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marco Antônio Pinheiro de Cristo**, **Professor do Magistério Superior**, em 11/02/2025, às 14:25, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **André Luiz da Costa Carvalho**, **Professor do Magistério Superior**, em 11/02/2025, às 18:32, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2428774** e o código CRC **714F5795**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Acknowledgment

Above all, I offer my deepest thanks to God, whose grace, strength, and guidance have been my foundation throughout this journey. His blessings have been my source of resilience and purpose, and without them, this achievement would not have been possible.

I am profoundly grateful for the invaluable support and encouragement of many individuals along the way. My sincerest appreciation goes to my esteemed research advisor, Professor Eulanda M. dos Santos. Her brilliance and unwavering guidance were instrumental throughout this journey. Her steadfast belief in my abilities motivated me to persevere through even the most challenging times, and witnessing her remarkable dedication as a scientist has been a constant source of inspiration. Without her mentorship and continuous support, this project would not have reached fruition.

I would also like to extend my gratitude to all the professors who have contributed to my academic journey. Their knowledge, insights, and commitment to excellence have greatly enriched my growth as a researcher and scholar.

I am eternally grateful to my family, especially in loving memory of my mother, Maria Olga de Queiroz Hermida. My heart is filled with immense gratitude for her love and enduring legacy. I am also deeply thankful for my beloved wife, Suely Ribeiro Sarges Hermida, and my cherished children, Paulo Cezar de Queiroz Hermida Filho and Juliana Ribeiro Sarges Hermida. My appreciation extends to all my brothers and sisters, whose unwavering support and encouragement have sustained me through this challenging journey, enabling me to bring it to a successful conclusion.

Resumo

Memes são uma forma popular de compartilhar mensagens com grandes audiências. Embora muitas vezes sejam engraçados ou informativos, memes também podem espalhar discurso de ódio e conteúdo prejudicial. Isso cria um desafio para as redes sociais, que precisam detectar e moderar esse tipo de conteúdo para tornar os espaços online mais seguros. Esta tese se concentra em melhorar a detecção de memes com conteúdo nocivo usando métodos avançados de aprendizado de máquina, incluindo modelos canônicos baseados em transformadores e os mais recentes modelos multimodais de linguagem de grande escala (MLLMs). O trabalho começa com uma revisão de literatura, destacando pontos fortes, fracos e desafios dos métodos existentes. Também é apresentada uma nova taxonomia para categorizar esses métodos, facilitando a comparação e a melhoria das técnicas de detecção. Na sequência, é proposto um método para melhorar o desempenho de modelos canônicos multimodais baseados em transformadores. Isso é feito ao adicionarmos um módulo chamado Bloco de Parâmetros Compactos nos codificadores dos transformadores. Os resultados dos nossos experimentos mostram que o método proposto supera diversos modelos mais complexos. A tese também explora o uso de modelos generativos, como os MLLMs, para detectar conteúdo agressivo em memes utilizando prompts específicos para orientar os modelos. Os resultados mostram que esses modelos conseguem identificar memes com conteúdo agressivo, mesmo não tendo sido projetados especificamente para essa tarefa. Porém, quando o nível de raciocínio multimodal exigido é muito elevado, o desempenho dos MLLMs reduz significativamente. Esta pesquisa contribui para a área ao oferecer melhorias práticas para os métodos atuais de detecção e ao explorar novas abordagens usando modelos generativos. Esses avanços são importantes para criar ambientes online mais seguros, respeitando a liberdade de expressão.

Abstract

Memes are a popular way of sharing messages with large audiences. They are often funny or informative, but memes can also spread hate speech and harmful content. This represents a challenge for social networks, which need to detect and moderate this type of content to provide a safer online environment. This thesis focuses on improving the detection of harmful content in memes using advanced machine learning methods, including canonical transformer-based models and recent multimodal large language models (MLLMs). First, the work presents a literature review that highlights the strengths, weaknesses, and challenges of current methods. It also introduces a new taxonomy in order to categorize current methods, making it easier for researchers to compare and improve detection techniques. Then, this work introduces a novel approach to enhance the performance of canonical multimodal transformer models. This is done by adding a specific module called Compact Parameter Blocks into the encoder segments of these models. The experimental results demonstrate that the proposed method outperforms several more complex approaches. The thesis also explores the use of generative models, such as MLLMs, to detect aggressive memes by using specific prompts to guide the models. The results show that these models can identify aggressive memes, despite not being explicitly designed for this task. However, the performance of MLLMs decreases significantly when the level of multimodality reasoning required is very high. This research contributes to the field by providing practical improvements to current detection methods and by exploring new approaches using generative models. These advances are important in creating safer online environments while respecting freedom of expression.

Table of Contents

List of Figures

List of Tables

1 Introduction

In the current digital age, global communication occurs at an unprecedented pace, impacting billions of people every day. The world population in 2024 is approximately 8.2 billion. Of this total, about 67.1% (5.5 billion) are Internet users, and approximately 63.7% (5.2 billion) are active on social media [1], as highlighted in Figure 1. In this virtual world, the spread of harmful online content has become a very important concern. This kind of content involves various forms, including hate speech, offensive language, harassment, misinformation, violence, and sexually explicit material [2]. Despite platform efforts to moderate this content, automated methods are crucial to managing the large amount of aggressive material that often goes undetected [3, 4].

Memes, in particular, represent a common way of spreading aggressive content. They have emerged as a popular form of communication, designed to quickly convey messages to large audiences. Although often humorous or informative, memes can also transmit hate speech and other forms of harmful content. In this context, social networks face the challenge of automatically detecting and moderating such content to ensure a safer and more inclusive online environment.

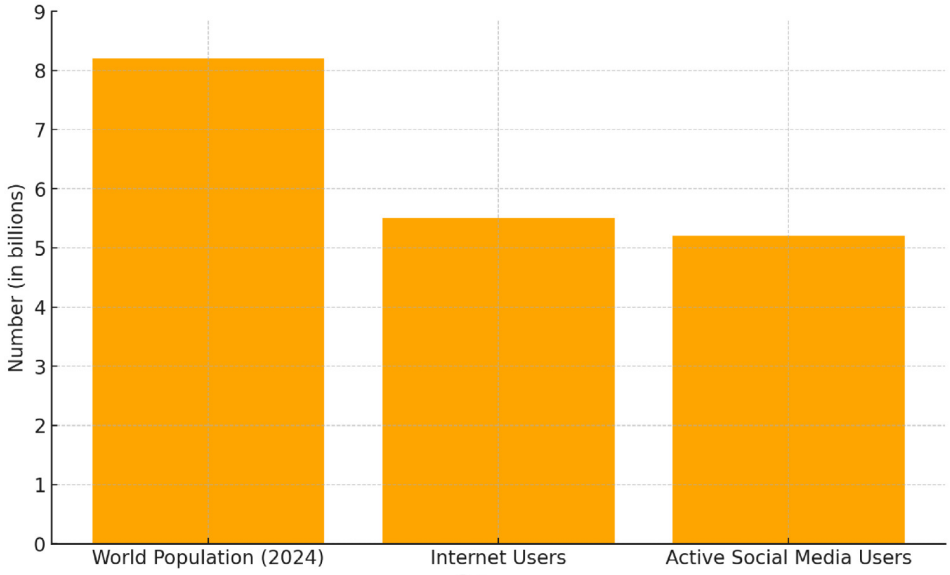

Figure 1 – Based on data from Statista (2024) [1], the graph provides a visual representation of global communication and social media statistics for 2024. Source: Author.

The detection of harmful content within memes is critical for online safety, especially because memes can propagate hate speech subtly or indirectly, through implicit meanings that are not easily detectable by conventional moderation systems [5]. Despite having different formats, such as short videos, GIFS, etc, memes are mostly multimodal, usually formed by images and text. While the meaning of a meme is highly dependent on context, their images and text often have no clear connection. Therefore, a thorough understanding of both modals is necessary to truly understand the message of the meme [6]. This is crucial to provide accurate tools that do not attack freedom of expression, which is a legitimate and inviolable right. As a result, identifying harmful memes requires robust machine learning models capable of effectively integrating multiple modalities.

In this context, the high performance rates reached by the transformer architecture in natural language processing (NLP) tasks [7] motivated many researchers to also employ these models to detect harmful memes in a multimodal approach. Currently, the most advanced methods for identifying harmful memes are based on the transformer architecture [8, 9, 10, 11, 12, 13]. These techniques use the transformer encoder component in multimodal models to gather both textual and image characteristics from memes, resulting in a comprehensive representation that arises from merging text and image elements.

More recently, multimodal large language models (MLLMs) have emerged as a promising innovation in natural language processing and computer vision. Models such as GPT-4V (Generative Pre-trained Transformer) [14], LLaVA (Large Language and Vision Assistant) [15], and Gemini [16] have demonstrated the ability to understand complex content by analyzing both text and images in an integrated manner. Therefore, MLLMs can also be used in the task of detecting aggressive content in memes, particularly at different levels of multimodal reasoning.

1.1 General objective

This thesis aims to enhance aggressive content detection in memes using multimodal machine learning models, first using canonical transformer-based multi-

modal approaches and then using MLLMs.

1.1.1 Specific objectives

The specific objectives are described in the following.

1. Conduct a structured review of the detection of hate speech in memes, identifying existing challenges and solutions, with a particular focus on the complexities of multimodal analysis (text and image). This includes proposing a new taxonomy for this topic and creating a foundational framework that can support future research.

2. Propose improvements to traditional multimodal transformers, identified in the previous review as the state-of-the-art architecture in this field. The proposed solution involves modifying the encoder blocks of these models by introducing compact parameter blocks to enhance their performance in detecting harmful content in memes.

3. Investigate the performance of MLLMs in identifying aggressive content in memes, assessing different levels of multimodal reasoning, and employing prompt engineering techniques to improve the models' overall performance.

4. Conduct a evaluation of the proposed methodology against state-of-the-art MLLMs. This involves assessing their performance in extracting and analyzing textual and visual features from memes, with an emphasis on detecting harmful content. The objective is to identify the strengths and limitations of both approaches in terms of accuracy, efficiency, and multimodal reasoning capabilities.

1.2 Thesis contributions

This thesis provides important contributions to the field of harmful content detection in memes, addressing the following key aspects:

(1) Comprehensive Review: A structured and thorough review of state-of-the-art research on hateful memes detection using machine learning is presented, summarizing and analyzing the approaches proposed in existing works.

(2) Taxonomy Development: A taxonomy for the detection of hateful memes through machine learning methods is introduced, facilitating the analysis of similar approaches and their respective results.

(3) Innovative Methodology: A novel method for enhancing the performance of multimodal transformer-based models in the detection of harmful memes is proposed. This method integrates a Compact Parameters Block (CPB) at the initial stage of the encoder, simplifying input data to enable the model to focus on critical multimodal information. This approach aims to improve the efficiency of encoding the attention mechanisms and the generalization capabilities.

(4) Generative Models Evaluation: The capabilities of leading generative models, including GPT-4V [14], Gemini [17], and LLaVA [18], are evaluated to detect aggressive content in memes. The strengths and limitations of these models, particularly in handling complex multimodal reasoning, are highlighted.

(5) Zero-Shot Prompt Analysis: The impact of well-designed zero-shot prompts is explored, demonstrating their ability to enhance classification accuracy in harmful meme detection.

These contributions are the core outcomes of this thesis and represent significant advances in meme analysis and harmful content detection. By addressing key challenges in this field, the research not only enhances existing detection methods but also introduces new perspectives on how multimodal and generative models can be leveraged for this task. The findings of this thesis provide a foundation for future studies, enabling researchers to develop more effective and robust techniques for identifying harmful memes.

1.3 Publications

- Hermida, Paulo Cezar de Q., and Eulanda M. dos Santos. "Detecting hate speech in memes: a review." Artificial Intelligence Review 56, no. 11 (2023): 12833-12851.
  https://doi.org/10.1007/s10462-023-10459-7

- Hermida, Paulo, and Eulanda M. Dos Santos. "Adding compact parameter blocks to multimodal transformers to detect harmful memes." Engineering Applications of Artificial Intelligence 137 (2024): 109136.
  https://doi.org/10.1016/j.engappai.2024.109136

- Hermida, Paulo Cezar de Q., and Eulanda M. dos Santos. "Exploring the Performance of Generative Models in Detecting Aggressive Content in Memes." Journal AI & SOCIETY - (under second peer review)

1.4 Thesis Organization

This thesis is organized as follows: Chapter 2 provides a comprehensive review of the theoretical foundation that supports the subsequent chapters. Chapter 3 presents a structured review of hate speech detection in memes, highlighting the challenges and existing solutions in the field. The content of this chapter was published in [19]. Then, Chapter 4 proposes the addition of compact parameter blocks to multimodal transformers to improve meme detection performance. This contribution of this thesis is published in [20]. In sequence, in Chapter 5 we evaluate the performance of MLLMs in detecting aggressive content in memes, analyzing different levels of multimodal reasoning and proposing improvements through prompt engineering. Finally, Chapter 6 concludes the thesis with a summary of key findings and suggestions for future research.

2 Background

This chapter presents theoretical aspects of the technologies used in this work. In Subsection 2.1, memes are defined, considering their format and content. In Subsection 2.2, the problem of detecting aggressive content in memes considering multimodal aspects is defined. In Subsection 2.3, the transformer architecture is described, which, based on the literature review we conducted, is the state of the art in detecting aggressive content in memes. Then, in Subsection 2.4, AutoEncoders are discussed as the foundation for the second main contribution of this work, which involves the use of compact parameter blocks in the encoder block of multimodal transformers to enhance performance. Next, in Subsection 2.5, the concept of MLLMs and generative models is introduced, linked to the third main contribution of this work, which evaluates the performance of these models in detecting aggressive content in memes. Finally, Subsection 2.7 covers topics related to prompt engineering.

2.1 The Evolution and Impact of Memes: From Humor to Harm

Memes have become a defining feature of online culture, acting as vessels for humor, social commentary, and shared experiences. The term “meme” was first introduced by Richard Dawkins in The Selfish Gene [21] to describe units of cultural information that spread through imitation. In the digital age, memes have evolved into multimodal forms of expression, often consisting of an image overlaid with text to create a blend of visual and verbal humor or meaning.

Although memes are often associated with positive and creative expression, their capacity for rapid dissemination has also been exploited for negative purposes. Aggressive content in memes usually targets individuals or groups, perpetuating stereotypes, hate speech, or misinformation. These memes can be subtle, cloaked in humor, or overtly offensive, aiming to harm the reputation, emotions, or safety

of others [22].

Groups frequently targeted by aggressive content often include marginalized communities, such as racial and ethnic minorities, women, and LGBTQ+ individuals. Such memes can perpetuate harmful social biases, fostering a hostile online environment that can lead to real-world consequences, including mental health challenges, discrimination, and even violence.

Below there is a concise summary of these groups, the impacts caused by aggressive memes, and some fundamental approaches for detection.

- **Minority Groups**
  - **Description**: Includes marginalized communities based on race, ethnicity, gender, sexual orientation, or religion.
  - **Impact**: Aggressive memes perpetuate stereotypes, incite hatred, or dehumanize individuals.
  - **Detection**:
    - Text analysis for discriminatory language.
    - Image recognition to detect offensive symbols or gestures.
    - Historical or cultural context analysis of the content [23].

Figure 2 shows some examples of memes in this category.

Figure 2 – A sample of aggressive memes targeting minorities. Extracted from FBHM dataset [24].

- **Teenagers and Young Adults**

- **Description**: This group heavily consumes memes on platforms such as TikTok, Instagram, and Reddit.
- **Impact**: May internalize aggressive behaviors or suffer from reduced self-esteem.
- **Detection**:
  - Identifying cyberbullying in online interactions.
  - Humor analysis to detect patterns of symbolic violence [25].

Figure 3 shows examples of memes this category.

Figure 3 – Examples of aggressive memes targeting Teenagers and Young Adult. Extracted from FBHM dataset [24].

- **Professionals and Institutions**
  - **Description**: Aggressive memes may target companies, governments, or public figures.
  - **Impact**: Reputation damage and security threats.
  - **Detection**:
    - Monitoring keywords related to institutions.
    - Image tracking to verify manipulation or editing of logos, public figures, etc. [26].

In Figure 4 it is shown a sample of memes from this category.

Figure 4 – Examples of memes focused on aggressive content against Professionals and Institutions. Extracted from MultiOFF dataset [27].

- **Women**
  - **Description**: Frequently targeted by misogynistic content in memes.
  - **Impact**: Reinforcement of gender inequality and online hostility.
  - **Detection**:
    - Identifying terms associated with hatred or objectification.
    - Tracking hashtags spreading attacks [28].

In Figure 5 there are some examples of memes from this category.

Figure 5 – Examples of aggressive memes targeting women. Extracted from Miso dataset [29].

- **LGBTQIA+ Communities**
  - **Description**: Often targeted by aggressive humor that seeks to delegitimize identities.
  - **Impact**: Promotion of hate speech and social exclusion.

– **Detection**:

* NLP to identify phrases or emojis with negative connotations.
* Social network analysis to map meme dissemination in specific groups [30].

In Figure 6 there are examples of aggressive memes in this category.

Figure 6 – Examples of aggressive memes against LGBTQIA+ Communities. Extracted from the FBHM dataset [24].

2.2 Problem Statement

Considering the multimodal aspect of memes, one way to express a generic multimodal representation X_m composed of n different modalities is summarized below:

$$X_m = f(x_1, ..., x_n) \tag{2.1}$$

In this work, the multimodal representation is composed of the following two modalities ($n = 2$):

1. Visual: $x_1 = \{I_1, ..., I_i\}$, where i is the meme index and I is the meme image.
2. Text: $x_2 = \{T_1, ..., T_i\}$, where T is the meme text.

The meme label is given by $Y = \{y_1, ..., y_i\}$. Here, $y \in \{0, 1\}$, x_1 and x_2 are the modality inputs, and the goal is to find the posterior probability $P(Y|x_1, x_2)$.

Taking into account these definitions, each modality x_i plays a different role in the general comprehension of the meme. The visual modality x_1 captures the features related to the image, including objects, colors, and spatial arrangements,

which often contribute significantly to the humor or message of the meme. In its turn, the modality of text x_2 involves analyzing the semantic content of the words, their syntax, and potential cultural references embedded in the text.

The function f, which combines these modalities into a single representation X_m, can be implemented using multimodal neural networks, such as transformers designed to process both images and text simultaneously. This approach aligns visual and textual information within a shared latent space, enabling a more comprehensive understanding of the meme's intent.

The ultimate goal is to predict the label Y, representing whether the content is aggressive or non-aggressive. By estimating the posterior probability $P(Y|x_1, x_2)$, the model determines the likelihood that the meme is classified as aggressive based on combined visual and textual inputs. This approach is essential for accurately capturing the complex and multimodal nature of memes, which often rely on the interplay between image and text to convey subtle or implicit meanings.

2.3 Transformer Architecture

Transformers [7] are a deep learning model designed to process sequential data. It relies heavily on the Self-Attention mechanism, which allows the model to focus on different parts of the input sequence when making predictions. Unlike traditional models, Transformers process input sequences in parallel rather than sequentially, making them more efficient. The architecture consists of two main components: Encoder and Decoder. The encoder processes the input sequence, generating a set of feature representations, while the decoder generates the output sequence based on the encoder representations. Each component is composed of multiple layers, each containing sub-layers for self-attention and feed-forward networks, along with layer normalization and residual connections to enhance learning stability. Transformers have become a foundational model for tasks such as language translation, text classification, and multimodal learning.

In this work, we focus only on the encoder block of transformers, as these components serve primarily as feature extractors. Consequently, they capture the es-

sential characteristics of the input data, enabling a more efficient representation of the multimodal information. Figure 7 shows the components of the encoder block of a canonical transformer.

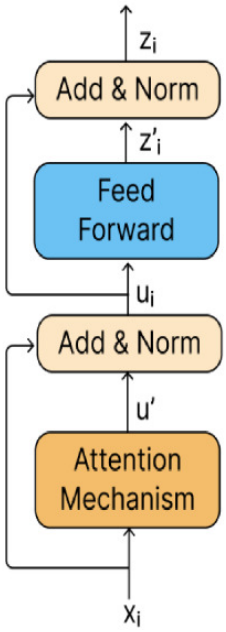

Figure 7 – Visual representation of the encoder block of the Transformers Architecture. Source: Author.

Given that x_i represents the embedding of the modality used as input to the initial encoder block, and d denotes the embedding dimension, the attention mechanism is defined using three equations, where $W_{h,q}, W_{h,k}, W_{h,v} \in R^{d \times d}$:

$$Q^{(h)}(x_i) = x_i W_{h,q}^T \tag{2.2}$$

$$K^{(h)}(x_i) = x_i W_{h,k}^T \tag{2.3}$$

$$V^{(h)}(x_i) = x_i W_{h,v}^T \tag{2.4}$$

In these equations, h refers to the specific head of attention, which varies from 1 to H, the total number of heads. The matrices $W_{h,q}$, $W_{h,k}$, and $W_{h,v}$ represent the weights for queries, keys, and values, respectively. The query matrix Q, defined in Equation 2.2, generates queries that identify relevant features of the input. The key matrix K, described in Equation 2.3, captures essential information from the input to serve as a reference. The value matrix V, shown in Equation 2.4, represents

content in specific positions to prioritize key information during the calculation of attention.

Equation 2.5 illustrates the calculation of attention weights $\alpha_{i,j}^{(h)}$, which determines the level of attention given by element x_i to another element x_j, within the context of head h. This mechanism enables each element of x to focus on other elements in the input sequence:

$$\alpha_{i,j}^{(h)} = softmax_j \left(\frac{\left\langle Q^{(h)}(x_i), K^{(h)}(x_j) \right\rangle}{\sqrt{d_k}} \right) \tag{2.5}$$

In Equation 2.6, the embeddings are combined using a weighted sum, using the computed attention weights. This produces a context vector that captures significant elements of the input sequence relevant to the current state of the model:

$$\mathbf{u}' = \sum_{h=1}^{H} W_{c,h}^T \sum_{j=1}^{n} \alpha_{i,j}^{(h)} V^{(h)}(x_j) \tag{2.6}$$

where $W_{c,h} \in R^{k \times d}$. The resulting output $\mathbf{u}_i'$ from the attention mechanism is combined with a residual connection and normalized, as represented in Equation 2.7:

$$\mathbf{u}_i = LayerNorm(x_i + \mathbf{u}') \tag{2.7}$$

Next, a feedforward layer is applied with a Rectified Linear Unit (ReLU) activation, as shown in Equation 2.8:

$$z_i' = ReLU(\mathbf{u}_i W_1^T) W_2^T \tag{2.8}$$

where $W_1 \in R^{d \times n}$ and $W_2 \in R^{m \times d}$. Finally, the output passes through another normalization layer, combined with a residual connection, as described in Equation 2.9:

$$z_i = LayerNorm(u_i + z_i') \tag{2.9}$$

Equations [2.6], [2.7], [2.8], and [2.9] mathematically define the structure of the original encoder block.

In summary, the encoder block, as depicted in Figure 7, processes input embeddings through multiple layers of self-attention, normalization, and feed-forward

operations. By combining these mechanisms, the encoder effectively captures and prioritizes relevant features of the input sequence, aligning them within a shared latent space. This iterative process refines the input representation, creating a robust feature set that can be used for further downstream tasks, as detecting aggressive content in memes. The defined equations [2.2] to [2.9] offer a clear mathematical framework for understanding the core functions of the encoder architecture, establishing the foundation for more complex multimodal models explored in this work.

In Chapter 4, we propose to integrate flexible Compact Parameter Blocks into the encoder segments of transformers. This approach presents similarities to the functioning of autoencoders, described in the next section.

2.4 Autoencoders

They are a type of neural network designed to learn efficient representations of data, typically for the purpose of dimensionality reduction, feature extraction, or data compression [31]. An autoencoder consists of two main parts [32]:

- Encoder: Maps the input x to a hidden representation h.
- Decoder: Reconstructs the original input from the hidden representation, producing $\hat{x}$.

Mathematically, these functions can be represented as:

$$Encoder : h = f_\theta(x) \tag{2.10}$$

$$Decoder : \hat{x} = g_\phi(h) \tag{2.11}$$

Where f_θ and g_ϕ are parameterized functions (typically neural networks) with parameters θ and ϕ.

The goal is to minimize the difference between the original input x and the reconstruction $\hat{x}$. A common loss function is the Mean Squared Error (MSE):

$$L(x, \hat{x}) = \|x - \hat{x}\|^2 \tag{2.12}$$

$$\min_{\theta,\phi} E_{x \sim p_{data}(x)}[L(x, \hat{x})] \tag{2.13}$$

where E denotes the expectation over the data distribution. This process allows the network to learn the most important features of the data while ignoring noise or less significant information [33]. Additionally, variants such as denoising autoencoders have been proposed to make the model more robust by intentionally corrupting the input data and training the network to recover the original data [34].

Autoencoders have proven to be highly useful in various applications, such as image denoising, anomaly detection, and data compression, due to their ability to capture complex structures in the data [35]. When combined with other models, they can serve as effective feature extractors, enhancing the performance of multimodal tasks [36].

Besides classical multimodal transformers, we also investigate in this thesis multimodal large language models. These models are described in the next section.

2.5 Multimodal Large Language Models and Generative Models

In the introduction we discussed the rapid progress of Large Language Models (LLMs) [37, 38, 39, 40, 41], and Large Vision Models (LVMs) [42, 43, 44, 45] in recent years, as shown in Figure 8. LLMs have shown impressive capabilities to understand text and perform complex reasoning, while LVMs excel in visual tasks, but often lack advanced reasoning skills. This complementary relationship between LLMs and LVMs has led to the Multimodal Large Language Models (MLLMs), which combine the strengths of both and other modalities to handle multimodal data, including text, images, audio, and video. MLLMs are based on massive LLM architectures enhanced with multimodal instruction tuning, allowing them to interpret images, generate website code from visual prompts, and understand complex

memes.

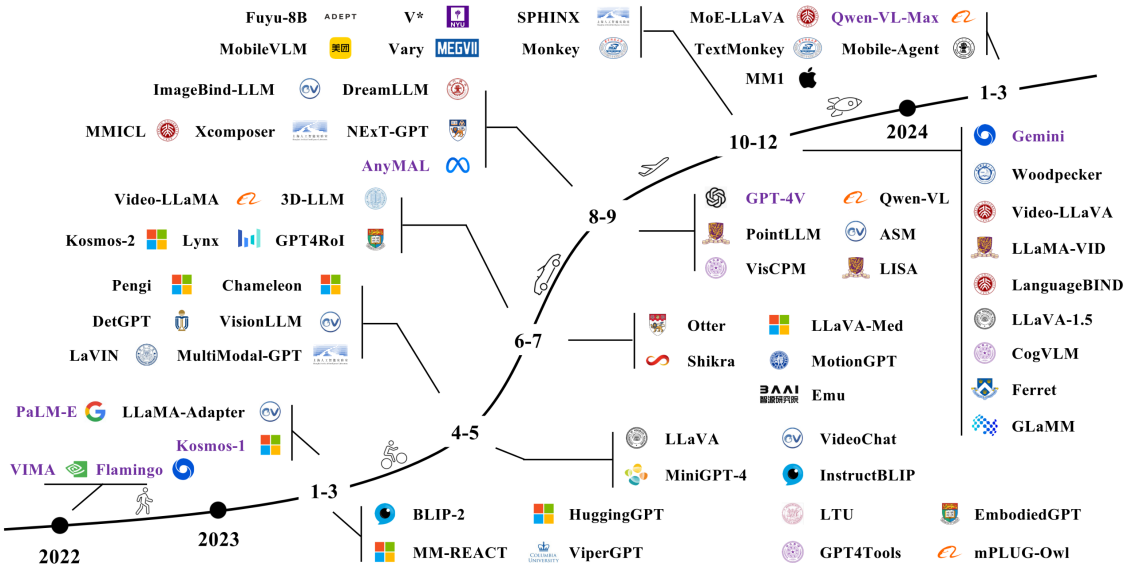

Figure 8 – The development of multimodal AI models from 2022 to early 2024, highlighting key models like GPT-4V, LLaVA, and Gemini. It illustrates the growing ability of these models to handle multiple types of data (like text, images, and video) and advancements in areas such as vision-language understanding and real-world applications. Source: [46].

MLLMs are advanced AI models capable of understanding and processing multiple modalities within a single framework. They are designed to integrate and reason across different types of data simultaneously. Figure 9 illustrates the architecture of a MLLM. The system receives inputs in different formats (e.g., images, audio, video, text). Non-text inputs pass through a Modality Encoder to transform them into a compatible representation (colored boxes), which is aligned with text data. After encoding, all modality-specific representations are unified through a Connector module, enabling the model to handle these diverse data types together.

The encoded and unified multimodal data is then fed into the LLM core. The LLM is responsible for processing and generating responses based on the integrated data, utilizing mechanisms such as Multi-Head Attention (MH-Attn) to attend to different parts of the input as needed. The LLM generates a text output or transforms the integrated data back into various formats (e.g., image, audio) via a Generator, allowing it to answer queries or produce results in multimodal formats.

The bottom of the figure shows further details of the model's internal processing. An MLP (Multi-Layer Perceptron) for initial transformations and a Q-Former with

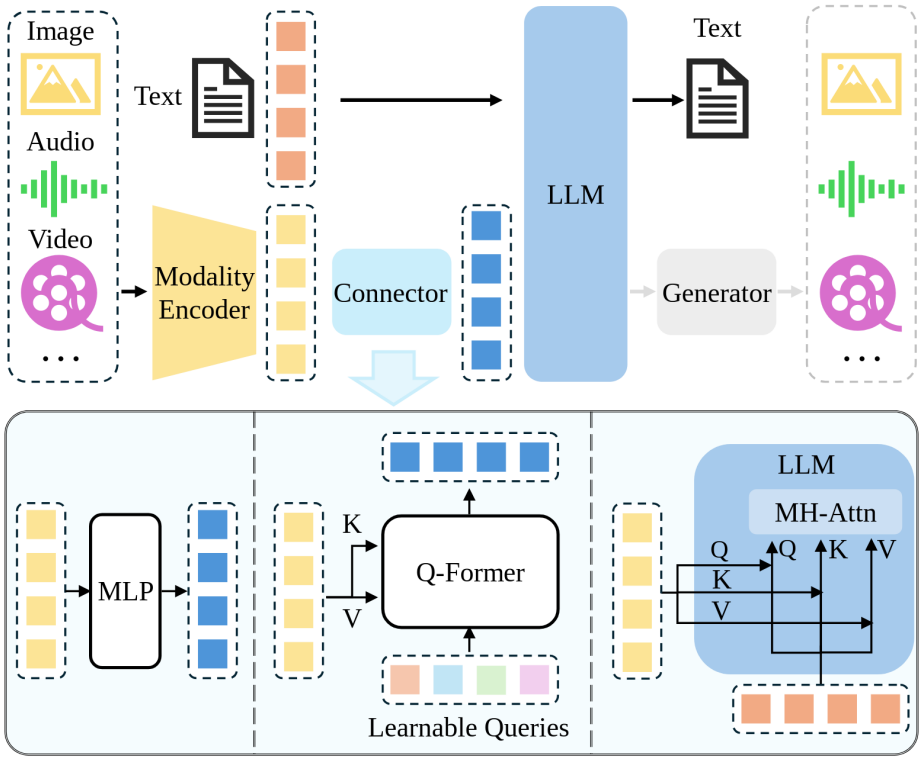

Figure 9 – Architecture of a MLLM: This diagram shows how multimodal inputs, including images, audio, video, and text, are processed. A Modality Encoder converts non-text data into compatible representations, which are unified by a Connector and processed by the MLLM using attention mechanisms. The system can generate text responses or multimodal outputs, integrating information across diverse data types for enhanced understanding and response generation. Source: [46].

Learnable Queries to structure the data into query-key-value (Q, K, V) pairs, which are then processed through attention mechanisms within the LLM. This architecture enables the LLM to understand and generate responses based on multimodal inputs, bridging various data types to enhance its reasoning and output capabilities across diverse applications.

In Chapter 5, we evaluate MLLMs for detecting aggressive memes. The models used are GPT-4V, LLaVA, and Gemini. In this context, we refer to these models as MLLMs because, in this work, we are not using them to generate new content. Instead, they are used to identify aggressive content in memes, which involves analyzing the relationship between text and images.

In contrast, generative models are designed to create new data reminiscent of the distribution of input data. Their primary focus is to generate new content, such as text, images, audio, or other formats, by learning patterns from training

data. While generative models can be multimodal (e.g., generating images from text prompts or vice versa), they are not necessarily limited to multimodal tasks and can operate within a single modality, such as generating only text.

MLLMs are commonly used for tasks that require simultaneous interpretation of language and visual data. For example, an MLLM can generate captions for images, answer questions based on image-text pairs, or identify specific objects in visual content based on textual prompts. In contrast, generative models are applied to generate new instances of data within one or more modalities. For example, a model like GPT-4V can generate text descriptions based on images, while DALL-E [47] can create images from text prompts, making it suitable for content creation and artistic generation.

It is well known in the literature that the main way to use MLLMs in new tasks is by performing Prompt Engineering, since traditional transfer learning techniques are often not possible. This involves creating specific instructions or questions for the models, helping them to give better answers or perform tasks more accurately, providing relevant and clearer answers. This aspect is discussed next.

2.6 Prompt

A prompt is the input provided to a language model or artificial intelligence system to elicit a specific response or output. It serves as a guide for the AI to understand the context, task, or question being posed. Prompts can be as simple as a question or as complex as structured instructions, depending on the desired outcome. For example, in natural language processing, a prompt might be a sentence fragment, a question, or a detailed scenario intended to generate a meaningful continuation or answer [48].

2.7 Prompt Engineering

Prompt engineering refers to the process of designing, refining, and optimizing prompts to improve the performance and accuracy of AI models, particularly large

language models. The goal is to craft prompts that maximize the model's ability to understand and respond correctly to a given task or query. This often involves iteratively experimenting with phrasing, structure, and context to achieve the desired output [49].

Prompt engineering has become a critical technique for leveraging the capabilities of large AI models in tasks such as text generation, question answering, and summarization. Prompt engineering for multimodal models involves designing and structuring inputs in a way that optimally aligns different data modalities to produce the desired output. Unlike traditional models that handle a single type of input, multimodal models require prompts that effectively integrate and guide multiple types of information simultaneously [50]. The challenge is to create prompts that not only capture the core content of each modality, but also employ their interplay to improve model understanding [51].

Figure 10 provides a comprehensive overview of various prompt engineering techniques designed to improve the performance of the language model in multiple dimensions. These techniques are organized into categories that address distinct aspects of model behavior and response generation, including handling new tasks, enhancing reasoning and logic, reducing hallucination, facilitating user interaction, optimizing performance, and more [52]. A brief description of each technique is presented below.

- **New Tasks Without Extensive Training**
  - **Zero-shot Prompting**: Enables the model to perform new tasks without needing specific training data for those tasks. Using its general knowledge, the model can generate responses based solely on instructions, even if it has not been explicitly trained on that specific task.
  - **Few-shot Prompting**: Improves the model's performance on new tasks by providing a few examples as context. This approach helps the model understand the task requirements with minimal examples, reducing the need for extensive training, and enabling adaptation to new tasks more effectively.

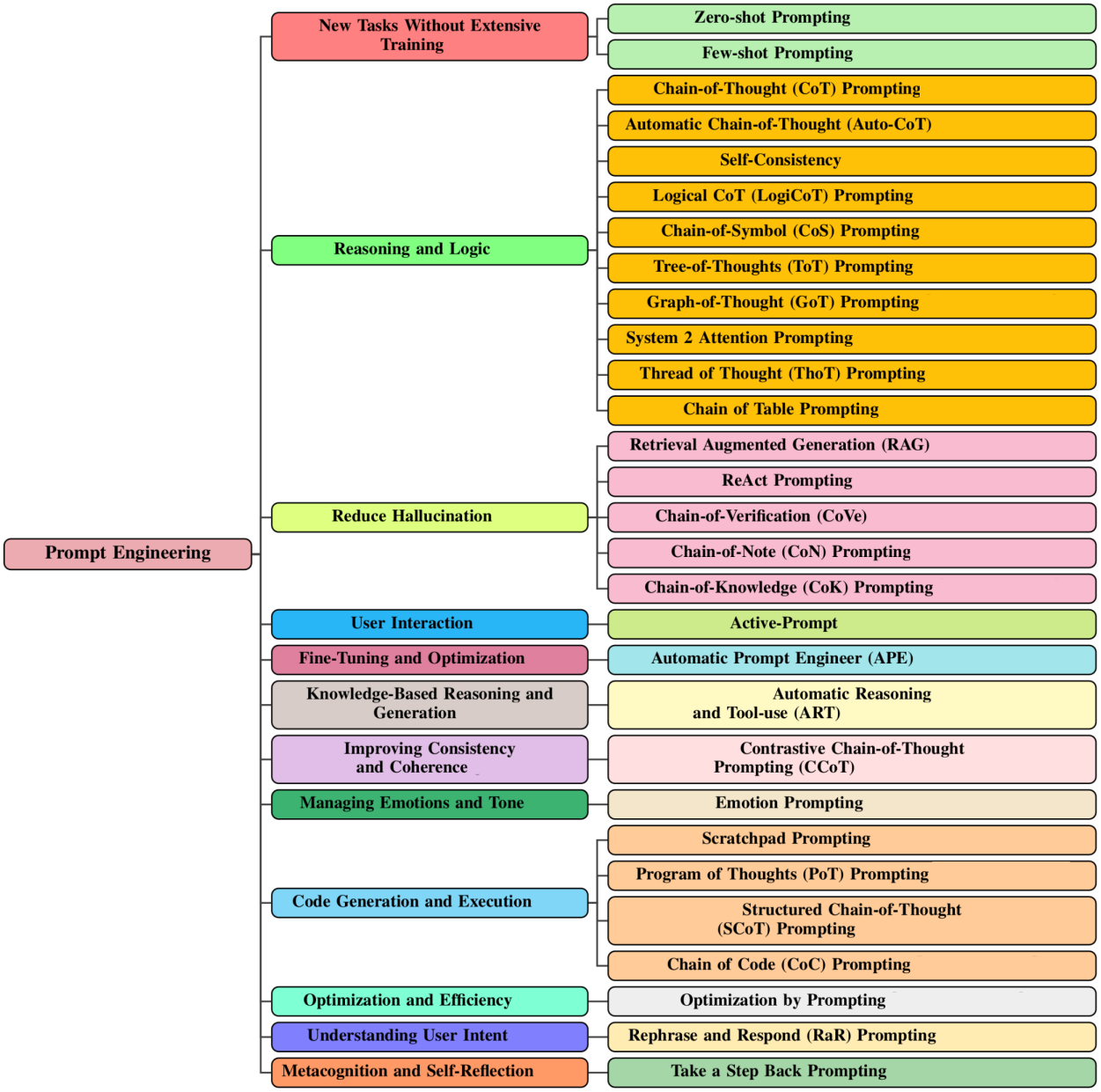

Figure 10 – Overview of prompt engineering techniques, organized by categories. Source: [52].

- **Reasoning and Logic**
  - **Chain-of-Thought (CoT) Prompting**: Guides the model to reason through a problem step by step, breaking down complex tasks into manageable steps. This technique improves logical reasoning by encouraging a systematic approach to problem solving.
  - **Automatic Chain-of-Thought (Auto-CoT) Prompting**: Automatically generates intermediate reasoning steps for the model to follow, further enhancing its ability to solve complex problems by structuring its approach.

- **Self-Consistency**: Encourages the model to generate multiple responses and evaluates them to choose the most consistent answer. This approach helps improve reliability by selecting the response that aligns best between different attempts.
- **Logical Chain-of-Thought (LogiCoT) Prompting**: It improves logical reasoning within chain-of-thought processes, helping the model maintain consistency and coherence in logical tasks.
- **Chain-of-Symbol (CoS) Prompting**: Uses symbolic reasoning techniques to handle tasks involving abstract or mathematical queries. By focusing on symbolic steps, the model can tackle problems that require manipulation of symbols or numbers.
- **Tree-of-Thoughts (ToT) Prompting**: Applies a tree-structured reasoning approach, allowing the model to explore multiple solution paths and choose the optimal solution. This approach mimics decision trees, allowing the consideration of alternative solutions.
- **Graph-of-Thoughts (GoT) Prompting**: Expands on the tree-of-thought approach by using graph structures, which represent relationships between different reasoning paths. This structure allows for more complex, interconnected reasoning.
- **System 2 Attention (S2A) Prompting**: Mimics human cognitive processes by emphasizing deeper, more deliberate reasoning. This technique encourages the model to reason with higher attention and thoroughness.
- **Thread of Thought (ThoT) Prompting**: Links various reasoning steps to create a cohesive, continuous thread of logic. This technique improves the ability of the model to maintain coherence in extended responses.
- **Chain-of-Table Prompting**: Organizes reasoning using a tabular structure to make complex information more manageable. This structured format helps the model break down information systematically for improved reasoning.

- **Reduce Hallucination**

  - **Retrieval Augmented Generation (RAG)**: Integrates external data sources into the model's responses, providing them with factual information and reducing the likelihood of hallucination.
  - **ReAct Prompting**: Combines reasoning with action, allowing the model to 'act' on information, such as verifying facts or retrieving data, before generating a response. This helps improve the accuracy of the responses.
  - **Chain-of-Verification (CoVe) Prompting**: Introduces a verification step in the response generation process, enabling the model to fact-check and confirm details before finalizing the output.
  - **Chain-of-Note (CoN) Prompting**: Encourages the model to take 'notes' throughout the response process, which can be cross-referenced for consistency and precision, reducing the likelihood of errors.
  - **Chain-of-Knowledge (CoK) Prompting**: Focuses on a knowledge-based reasoning process, prompting the model to verify facts by referencing internal knowledge, thus enhancing the factual accuracy of responses.

- **User Interaction**
  - **Active Prompting**: Adapts prompts dynamically based on user input or model responses, creating a more interactive experience. This technique allows the model to adjust its responses in real time, making interactions more responsive and contextually relevant.

- **Fine-Tuning and Optimization**
  - **Automatic Prompt Engineer (APE)**: It uses automated methods to optimize prompts for specific tasks, improving model performance by refining prompt structure and content to produce better responses for particular use cases.

- **Knowledge-Based Reasoning and Generation**
  - **Automatic Reasoning and Tool-use (ART)**: Allows the model to incorporate specialized reasoning techniques and use external tools for task-

specific accuracy, enhancing the quality of responses in specialized areas like scientific or technical domains.

- **Improving Consistency and Coherence**
  - **Contrastive Chain-of-Thought (CCoT) Prompting**: Prompts the model to evaluate multiple response options and select the most coherent and consistent answer. This technique improves reliability by reducing the variability in responses.
- **Managing Emotions and Tone**
  - **Emotion Prompting**: Provides control over the emotional tone of responses, enabling the model to match the desired tone, whether formal, empathetic, or neutral, according to user expectations or context.
- **Code Generation and Execution**
  - **Scratchpad Prompting**: Encourages the model to maintain a 'scratchpad' of intermediate steps or notes, which is particularly helpful in generating and debugging code, as it allows the model to build solutions iteratively.
  - **Program of Thoughts (PoT) Prompting**: Guides the model through a structured, step-by-step approach to coding tasks, ensuring logical progression and improving accuracy in complex programming scenarios.
  - **Structured Chain-of-Thought (SCoT) Prompting**: Uses a highly structured approach for code generation, breaking down tasks into logical components to ensure code accuracy and functionality.
  - **Chain-of-Code (CoC) Prompting**: Break down coding tasks into sequential steps, allowing the model to generate code in logical segments, thus improving both accuracy and readability.
- **Optimization and Efficiency**

  - **Optimization by Prompting (OPRO)**: It helps to simplify model responses and operations, enhancing computational efficiency by reducing unnecessary processing and focusing on essential information.

- **Understanding User Intent**

  - **Rephrase and Respond (RaR) Prompting**: Helps the model clarify and rephrase user input, improving comprehension of user intent and generating more accurate responses by aligning closely with user needs.

- **Metacognition and Self-Reflection**

  - **Take a Step Back Prompting**: Encourages the model to pause and review its response, promoting self-reflection, and enhancing reliability by allowing the model to reassess and correct potential errors in its output.

In conclusion, the prompt engineering techniques outlined in this section represent a diverse toolkit for enhancing language model performance. By categorizing methods into areas such as task generalization, reasoning enhancement, hallucination reduction, user interaction, and optimization, we gain a structured understanding of how these techniques address specific challenges in NLP. Each technique is designed to target particular aspects of model behavior, from improving the factual accuracy to refining logical consistency, emotional tone, and user intent comprehension.

2.8 Final Remarks

In this chapter, we detail the theoretical foundation that will be used at various points in the next chapters of this thesis. We begin with a brief overview of memes and the various aspects related to their use in propagating aggressive content. We define the primary problem addressed in this thesis: the automatic detection of aggressive content in memes. Following this, we discuss transformer architectures, which currently represent the state of the art in this field. We introduce the concepts of Autoencoders and Multimodal Large Language Models, which, together

with prompt engineering, form the set of tools utilized throughout this work. In the next chapter, we will explore our first main contribution in detail: a review of the literature aimed at identifying studies addressing this topic.

3 Detecting Hate Speech in Memes: a Review

In this chapter, we present the first main contribution of this thesis, which is a survey discussing several recent researches aimed at detecting hate speech in memes. We list the most recent research, synthesize and discuss the approaches proposed in the current literature by providing a critical analysis of these methods, highlighting their strengths and points to improve. We also introduce a taxonomy to allow grouping similar approaches. This survey is published in [19].

3.1 Introduction

Memes represent a possible source for spreading hate speech through social networks. They have gradually adapted to the internet format and focus on quickly conveying a message to as many people as possible. Their message can be positive or funny, but can also carry hate speech toward specific groups within our society. The hateful aspect can be directly embodied in memes message, but it can also be indirect by endorsing hateful speech when propagating their message on the internet. However, developing machine learning-based models to point out whether or not a meme contains hate speech is highly challenging, especially due to two reasons. First, hate speech is not clearly defined. For instance, some definitions can be found in [53, 54, 55, 56, 57, 58]. A hateful meme may contain personal attack, racial abuse, attack on minority, among others. Therefore, this non-standard definition may limit the entire machine learning process on detecting new memes with hate speech, even the annotation of datasets is challenging.

The second reason is that memes are mostly multimodal. For instance, the memes investigated in all works surveyed in this chapter are formed by images and text [6]. In this case, the analysis of the content of a meme must take into account both modalities to allow capturing the original meaning of the meme's content. If analyzed separately, image and text may have no relation to the original meaning.

These aspects should be considered while designing a method for detecting hateful memes, since the task involved is: given an image and the text superimposed on that image, detect whether this set (image and text) take on hateful meaning or not.

The research results summarized in Figure 11 show that there are several works providing survey on methods that can detect hate speech. Among these, only the work in [59] addresses this issue considering memes as propagators of hateful content. However, such paper does not describe any work that effectively tackles hateful memes detection, since their focus is driven toward the Visual-Linguistic domain. In this context, the authors assigned the task of detecting hate speech in memes as an application problem in the Visual-Linguistic domain. Consequently, they do not survey innovative methods for hateful memes detection specifically.

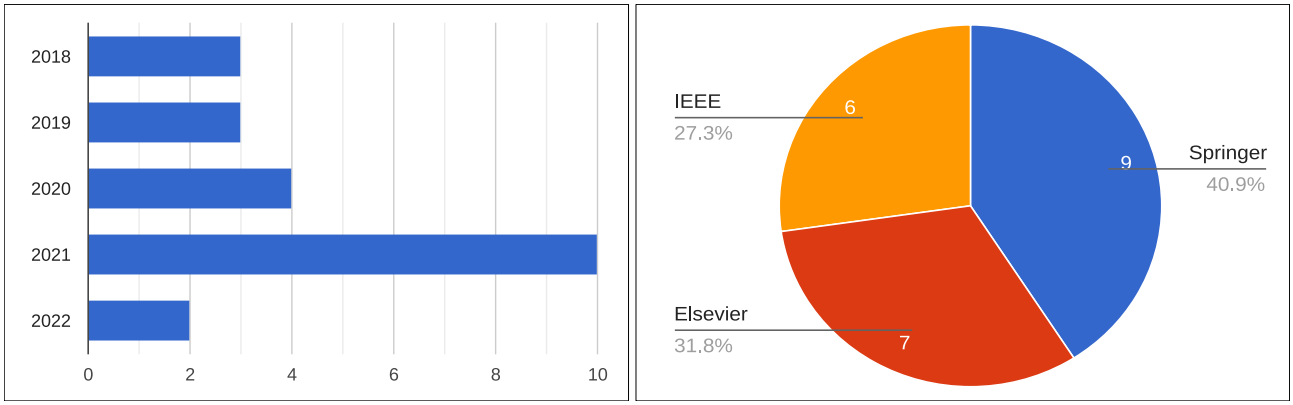

Figure 11 – Timeline of surveys publications in hate speech identification. Timeline showing a significant increase in a few years. Source: Author.

Considering this scenario, the objective of this chapter is to try to fill this gap in the literature by identifying the most recent works that address hateful memes detection. The main contributions of the this review can be outlined as follows.

- A structured and comprehensive review of the state-of-the-art research in hateful memes detection using machine learning is presented. We summarize and discuss the approaches proposed in the reviewed works.
- A taxonomy for hateful memes detection using machine learning methods has been put forth, allowing us to analyze similar approaches and their achieved results.

- A thorough critical analysis of the methods currently available, showing their strengths and points to be improved, is provided.
- An analysis of the evolution of the research domain to explore the open and trending research challenges of the hateful memes detection is presented.

3.2 Proposed Taxonomy

In the previous section, we mentioned that the analysis of the content of a meme must considerate the fact that memes are multimodal data. Therefore, in this section we analyze only works proposing multimodal approaches. We propose to categorize these works according to three levels of features, as depicted in the graphical taxonomy shown in Figure 12.

The first level involves the use of attention mechanisms, providing two main categories: 1) non-attention mechanism-based methods; and 2) attention mechanism-based methods. Here, the attention mechanism aspect used to define this level in our taxonomy considers the perspective of generating multimodal representations for the memes, i.e. attention-based methods that learn joint representations of multimodal content. Therefore, works employing models composed by attention mechanism as feature extractors do not fit into this group. This category is more related to how attention mechanism is used rather than to its presence in the solution.

In the second level of our categorization, we consider that, broadly, most of the current research work on hateful memes detection can be grouped taking into account how text and image from memes are tackled. In this case, two concepts are introduced here:

- **Restricted**: These are works that use the information present in the image and text from the meme directly, using feature extractors for the text and the image.
- **Extended**: These are works that use additional data to enforce multimodality. The additional data are present in the image and meme text but are indirectly

extracted. Tags of objects, age, gender, emotion, and sentiment analysis are examples of additional data.

Considering the second level, methods from the attention mechanism-based group are divided into restricted and extended. In terms of non-attention mechanism-based methods, all the papers from this group cited in this review fall only under the restricted category. We may point out at least two main reasons for this domain of restricted methods. The first reason is related to the models' architectures, since using extended data without attention mechanism requires more complex architectures because the additional information must be extracted from the text and image of the meme and then properly combined to generate the expected result. The second reason refers to the reduced scope of the datasets investigated. Works included in the non-attention mechanism-based group are mostly focused on specific categories of memes, such as memes with sexist messages and memes related to the American elections. In this case, using the complementary information provided by external data is expected to be less required. A different scenario is observed in the attention mechanism-based group, where there is inherent benefit to performance by using external information in conjunction with the information present in the image and text from the meme directly due to the very broad range of categories of memes investigated by works from this category.

Finally, the third level in the proposed taxonomy takes into account the feature extraction process performed, which is achieved by auto-feature extraction and hand-crafted techniques.

- **Auto-feature Extraction**: Refers to techniques that automatically extract features, such as representation learning or deep learning-based methods. These approaches allow models to learn and identify the most relevant features from the data without direct human intervention.

- **Hand-crafted technique**: Refers to feature extraction methods manually designed by experts. In these techniques, features are carefully selected and defined based on prior knowledge of the data and the problem domain, requiring a deep understanding of what may be relevant for the given task.

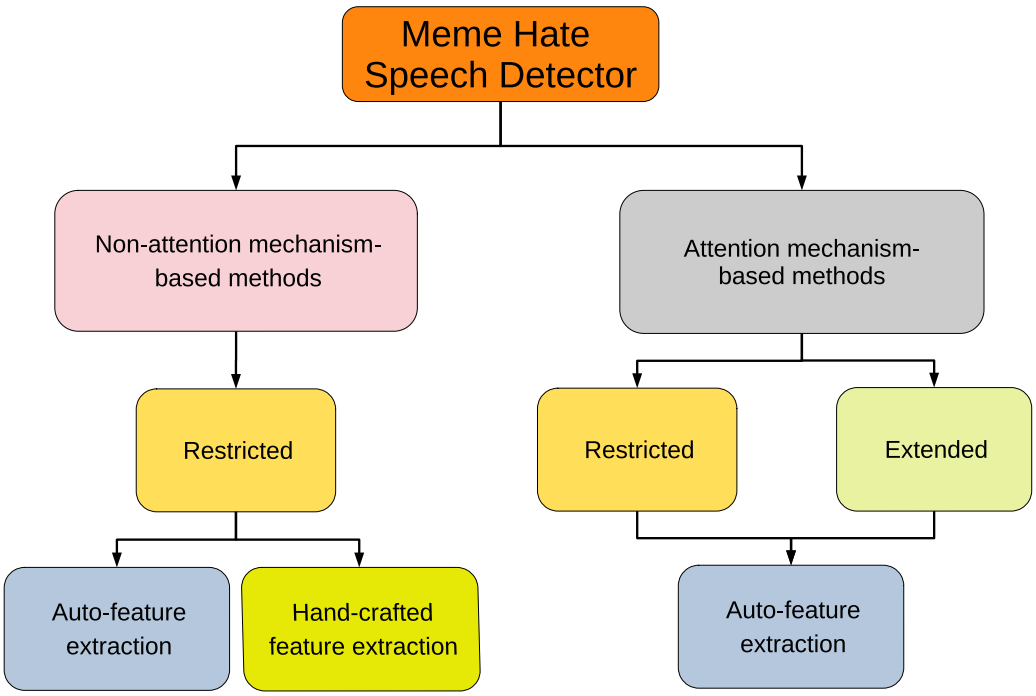

Figure 12 – Proposed taxonomy considering three levels of features. Source: Author.

In terms of the attention mechanism group, all works are clearly grouped into the auto-feature extraction category because their methods use the Transformer architecture.

It is important to mention that, given that all studied methods are multimodal learning models, an obvious category would be defined by the data fusion approach employed. However, despite having different fusion strategies available in the relevant current literature on data fusion, these strategies do not present specific features known to be strongly associated with specific groups of approaches dealing with hateful meme detection. Consequently, fusion strategies do not differentiate these approaches because they are widely employed in all groups of methods. On the other hand, the description of the methods reviewed in this work includes the fusion approach, since the representativeness of the space resulting from fusion is fundamental to successfully perform meme classification. Due to this reason, here we define and describe the three main different fusion strategies, according to [60].

- **Early Fusion**: It extracts feature vectors (f) from individual modalities i, combining them in a fusion module (FF) by concatenating or pooling, for instance. Then, it employs an interpretation unit (IU) to make a final decision (D). This

strategy is also called feature level fusion. Figure 13 illustrates a generic early fusion model.

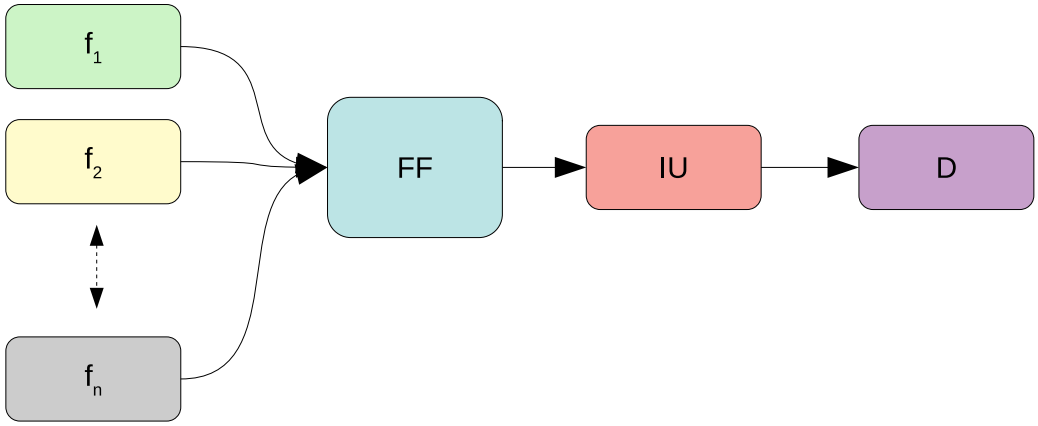

Figure 13 – A generic early fusion model. Each feature vector (f_i) is concatenated in a fusion module (FF) and the result of the fusion is processed in the interpretation unit (IU), responsible for generating a final decision (D). Source: Author.

- **Late Fusion**: In this approach, also known as decision-level fusion, features are also extracted from each modality i. Here, however, each feature vector is used to feed one (IU) per modality. Thus, each (IU_i) assigns indiviual decisions (D_i) to each input instance. After that, the individual predictions are grouped in the Decision Fusion (DF) phase using an aggregation function, e.g. averaging, majority voting, weighted voting, etc. Finally, another (IU) processes the result to provide the final decision (D). Figure 14 summarizes a generic late fusion model.

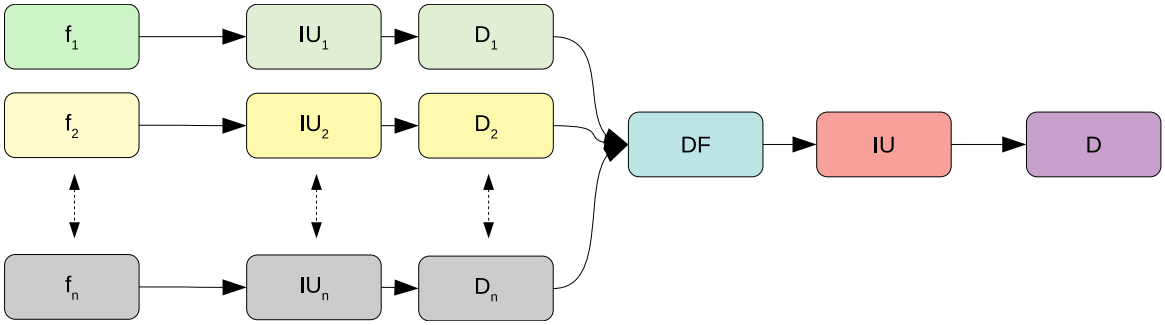

Figure 14 – A generic late fusion model. Each feature vector (f_i) is processed in the interpretation unit (IU_i), generating a partial decision (D_i). The decisions feed the decision fusion module (DF), whose result is processed by another interpretation unit (IU), which generates the final decision (D). Source: Author.

- **Hybrid Fusion**: It includes early and late fusion at the same time. Figure 15 illustrates a generic hybrid fusion model.

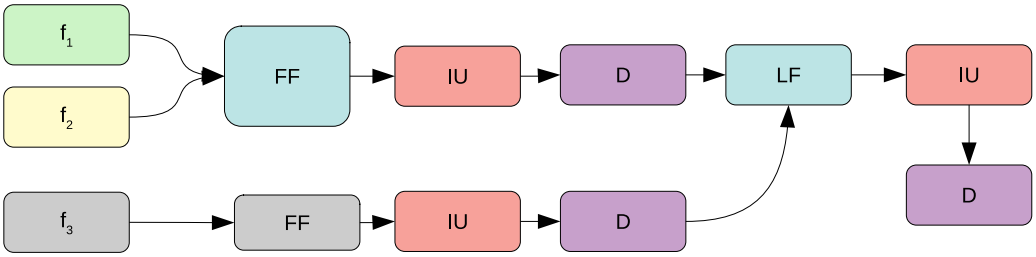

Figure 15 – A generic hybrid fusion model. The early fusion merges with a later fusion in a last merge module (LF), whose result is processed by an interpretation unit (IU), which generates the final decision (D). Source: Author.

In the next section, each of the nodes of the proposed taxonomy will be explored and all reviewed works will be detailed and discussed.

3.3 Existing Approaches to Detect Hateful Memes

This section provides an overview of the state-of-the-art of detecting hate speech in memes. For this, the taxonomy proposed in Section 3.2 is used. In Subsection 3.3.1, we present and discuss approaches that rely on non-attention mechanism, whereas in Subsection 3.3.2, approaches that use attention mechanisms are addressed. Subdivisions of these subsections are also performed according to the taxonomy.

3.3.1 Non-attention Mechanism-based Approaches

This section describes works that do not use attention mechanism to learn joint representations of the multimodal data. Here, it is important to consider that some works in this category employ methods which present attention components, such as Transformers. However, these methods are used as feature extractors, usually to obtain textual features. In this case, textual and visual features are first obtained and then combined using a traditional fusion approach, especially early fusion. In the second level of the proposed taxonomy, these works are all grouped

into the restricted category. This means that only features directly extracted from the meme's image and text are used to represent the memes to the classifier. These works are further subdivided, depending on the feature extraction method used, as described in the next subsections.

3.3.1.1 Hand-crafted Feature Extraction-based Methods

There are methods in the literature focused on discovering, understanding, characterizing, and improving features that can be handcrafted from the text and the image of memes. The handcrafted features are mainly combined with traditional supervised learning algorithms, such as support vector machines (SVM), decision trees (DT) and k-nearest neighbors (kNN), to detect hateful memes.

In [61], the authors employ text and image handcrafted features specifically developed for detecting sexist memes using mono and multimodal approaches. They used a dataset composed of 800 memes with sexist/non sexist content depending on visual and/or textual aspects. The memes from this dataset were collected from social media platforms, such as Facebook, Twitter, Instagram and Reddit. The textual features are obtained by Bag-of-word [62], providing a 2048-dimensional feature vector. In terms of visual features, 23 simple techniques are used [63]: Coarseness, Contrast, Differentiability, Line likeness, Roughness, Edge density, Entropy and Measure of Enhancement, Local Binary Pattern, Histogram of Oriented Gradients (HoG)), Chroma variance, Number of regions, Color, Color histogram in HSV (hue, saturation, value) color space, Mean and Standard deviation of colors in RGB color space, Auto-correlogram, Resource congestion and Sub band entropy, Image complexity, Measure of the degree of focus, Vector of aesthetic characteristic, Percentage of skin measure and Number of faces. These extractors generated a 4418-dimensional feature vector.

Four traditional supervised learning algorithms are investigated: SVM, DT, 1NN and Naive Bayes (NB). In addition, the authors compare both early [64] and late data fusion approaches. In the first, the textual and visual feature vectors are concatenated. In this approach, DT reached the best result, precisely accuracy: 0.696, Recall: 0.696 and F1 score: 0.696. In the second approach, SVM obtained

the best result, which outperformed the early fusion results. The attained rates are accuracy: 0.759, recall: 0.760 and F1 score: 0.759.

However, when comparing the multimodal results to the unimodal ones, SVM using only the textual features reached rates similar to the best performing multimodal model. The authors concluded that the high number of visual features contributed negatively to the result of the multimodal approach. Another aspect is the fact that the visual feature vector was more than twice higher dimensional than the textual feature vector, making it difficult to carry out the fusion strategy in a way that the textual features could compensate errors caused by the visual features. Finally, the dataset used to conduct the experiments was composed by 800 memes collected from social media and labeled by volunteers. Since no guidance was provided to the volunteers to avoid influencing their judgment, labels may be biased, thereby affecting the result reached.

It is important to note that the choice of the feature extraction method is fundamental for the success of the learning models and directly affects the results. The next category of methods try to avoid this drawback using models that learn features directly from the raw inputs, as described in the next section.

3.3.1.2 Auto-feature Extraction-based Methods

The process performed in this group of methods is very similar to the process observed in the previous category. The difference is the feature extraction step, which is conducted by deep learning models whose deep layers act as a set of feature extractors. Hence, the set of features is learned directly from observations of the input data.

The work presented in [27] shows different combinations of deep learning models employed to extract features from text and image aiming at detecting offensive memes. For the textual features, GloVe is first used to obtain vector representations for the words. GloVe is an unsupervised learning algorithm trained to aggregate global word-word co-occurrence statistics providing word embeddings. Then, three different approaches are individually employed to extract textual features: Long Short Term Memory (LSTM) [65]; Bidirectional LSTM (BiLSTM) [66];

and Convolutional Neural Network (CNN). In the first, two stacked LSTMs are used for feature extraction. In the second, only one BiLSTM is applied. Finally, in the third approach, three convolutional blocks (convolutional layer + maxpooling layer) compose the CNN responsible for feature extraction. In terms of visual features, only the CNN VGG-16 [67] is the feature extractor.

The different feature extractors are used to compose three multimodal approaches: Stacked LSTM + VGG16; BiLSTM + VGG16; and CNNText + VGG16. In the two first models, visual and textual feature vectors are concatenated before being sent to the classification layer (a Neural Network), following the early fusion approach. The CNNText + VGG16 model additionally feeds the final concatenated feature vector to a stacked LSTM model, whose output is combined with the visual features to represent the meme to the classification layer. Besides the three multimodal approaches, they conducted experiments with seven unimodal approaches: Logistic Regression [68], NB [69], Stacked LSTM [65], BiLSTM [66], and CNNText employing text features; and VGG16 to perform offensive meme classification solely based on visual features.

The authors have created a dataset composed with only 743 memes related to the 2016 United States presidential election to conduct their experiments. The results indicate that both groups of methods reached very low accuracy, being the logistic regression model using only text features the highest performing approach (accuracy = 0.58). According to the authors, the results might be influenced by the dataset itself due to biases caused by annotators, small number of samples and unrepresentative data. Although deep learning feature extractors are generally expected to provide representative features, and working in a multimodal fashion is the ideal solution when dealing with offensive meme detection, this was not the case in this paper. Indeed, the authors indicate that human moderation is still necessary.

In [70], textual and visual features are also obtained separately. First, the text is isolated from the image using the Tesseract Optical Character Recognition (OCR) method. Second, visual and textual features are extracted. The Bidirectional Encoder Representations from Transformers (BERT) [71] is the model used to extract the features from the text, providing a 768-dimensional feature vector. In

turn, the authors employ VGG-16 to extract the visual features, returning a 4096-dimensional feature vector. Then, the two feature vectors are concatenated in an early fusion fashion. The combined 4864-dimensional feature vector is finally used as input to a fully connected Neural Network, which is responsible for providing a score indicating the level of hate speech in the meme.

The authors in [70] have compared their multimodal approach with its unimodal counterparts on a dataset composed by hateful memes collected using a search engine. The dataset contains 5020 memes, being 1695 hate memes of three categories: 643 with racist content; 551 related to Jews; and 501 to Muslims. The remaining 3325 instances are memes without hate speech, which were obtained from the Reddit Memes [72]. The comparison results show equivalent accuracy rates attained by the multimodal (0.833) and the unimodal (0.830) approaches. In the latter, the reported accuracy is obtained using the visual features since this unimodal model outperformed the one trained using only textual features. The authors conclude that the multimodal approach can filter some memes distributed on social networks, but the moderation of a human being is still necessary in many cases. They have also observed that their model can be used to indicate which image and text combinations provide the highest level of hate speech, allowing the detection of new hate speech memes.

The comparable results in terms of accuracy reached by the multi and unimodal approaches can be explained by two reasons. First, the annotation process of the hateful memes class may compromise the dataset due to biases inherent in the search engine used to collect the data. Another challenge is the difference between the dimensionality of visual and textual feature vectors. The visual feature vector is much higher dimensional than the textual feature vector. This may also explain the superior performance shown by the unimodal approach trained with only visual features.

The work presented in [73] reached conclusions different from those achieved in the previous works. They propose a method to detect hate speech in memes that extracts textual features using one model, precisely VGCN-BERT [74]; and three CNNs to extract visual features: ResNet50, ResNet152, and VGG-16. Again, early fusion is also performed by concatenating in pairs the text embedding with

each image embedding to create different multimodal models. When comparing the multimodal models to the unimodal approaches in a dataset created with 1600 memes related to Italian political affairs, the results showed that the best option was the combination VGCN-BERT + ResNet50, achieving 81.69 of AUCROC.

Despite the multimodal outperformed the unimodal approach, the authors report as problems the presence of concise text in some memes and some subjectivity due to the dataset's annotators. They also suggest using other architectures to extract visual features, such as VGG-19 [67], EfficientNet [75], and multilingual neural networks, like mBERT [76] and XLM-RoBERTa [77]. Moreover, the dataset investigated contains memes specific to Italy and focused on political content.

One major challenge to works reviewed thus far is that the proposed methods perform rather poorly compared to human performance. Additionally, these works do not explore all the potential of multimodality in memes, since multimodal approaches were less performing than unimodal models, indicating that a more refined understanding of multimodality is necessary. Works in the next category of methods focus on studying whether attention mechanism-based models have the potential to improve both feature extraction and multimodal fusion.

3.3.2 Attention Mechanism-based Approaches

This section describes works whose approaches use attention mechanism to learn joint representations of multimodal content. Following the proposed taxonomy, these works are subdivided into two groups. The first groups methods that follow the restricted approach, i.e. they employ information present directly in the image and text of the memes obtained using feature extractors. The second group involves extended approach-based methods, which use additional information extracted indirectly from the image and/or the text of the meme. In terms of the third level of the taxonomy, all works discussed in this section fall into the auto-feature extraction category.

It is worthy nothing that the attention-based methods reviewed in this chapter are all designed according to the Transformer architecture, which is composed by

Encoder and Decoder layers. The Encoders process the input data and allocate more weight to the parts that are more relevant. Such an information is passed to the next Encoder layers, creating a more accurate representation of the input data. The data obtained as output of the last Encoder are fed into the first Decoder layer, which does the reverse process to generate the output data. The relative importance of the Encoder block output, which are the input data to the Decoder, is also taken into account.

3.3.2.1 Restricted Methods

It is important to observe that all works previously mentioned in this chapter deal with challenges caused by the datasets, which might be insufficient to widely represent the problem and might not allow the multimodal information between the two modals to be fully explored and used. In order to contribute to this matter, the Hateful Memes Challenge published a dataset composed by a training set of 8500 images, a validation set of 500 images and a test set of 1000 images. The dataset has 12 categories of hate speech: comparison to animal, comparison to object, comparison with criminals, exclusion, expressing disgust/contempt, mental/physical inferiority, mocking disability, mocking hate crime, negative stereotypes, use of slur, violent speech and other.

To encourage the proposal of methods capable of truly understanding multimodality, the Hateful Memes Dataset (HMDC) [24] provides text and vision confounders. These confounders lead to changing a hateful meme to a not-hateful one, or vice-versa, by swapping either image or text only. In Figure 16 we highlight this characteristic. The meme in the first column is labeled as hateful, while in the second and in the third column we see its confounders whose image or text allows flipping its label to a not-hateful meme. Therefore, a model must be able to tackle multimodal reasoning so as to classify the original meme and its confounders correctly. Due to its challenging nature, this is the dataset employed in the next works discussed in this chapter.

In [78], the authors first added 328 more memes to the HMDC dataset. The additional instances were obtained from the Memotion dataset [79]. Their approach

Figure 16 – Example of confounders memes. Image on the left side shows a hateful meme, while images on the right side and middle show its confounders resulting in flipping its label to a not-hateful meme. Source: [24].

involves feeding VisualBERT directly with text tokens. For the visual features, a ResNeXT-152-based Mask-RCNN model is employed, providing 100 regions from the main meme image, and a 2048-dimensional feature vector from each region. In the sequence, the visual embeddings are projected into the textual embedding, which is then fed to the transformer layers. The VisualBERT model employed was pretrained on the Conceptual Captions (CC) [80] dataset. An ensemble composed with 27 base models was obtained by making changes to the values of the VisualBERT model hyperparameters. Then, the late fusion approach via majority voting as fusion function was applied to predict a class to the memes. The authors attained 0.76 as accuracy rate and 81.08 for the AUCROC metric, outperforming several baselines. On the other hand, this approach is affected by problems caused by object detectors, which may failure on finding objects in the image.

The work discussed in this section shows results reinforcing that working on solutions that detect hate speech in memes using attention mechanisms seems like a promising idea. The ensemble approach employed in [78] was also shown to reach higher performance than single classifier-based multimodal methods. Works discussed in the next section employ features indirectly extracted from the memes. We focus on analyzing whether or not this approach helps to improve results.

3.3.2.2 Extended Methods

The group of works reviewed in this section relies on the hypothesis that it is necessary to take into account more than the image and text description of memes to

successfully detect hateful memes. In order to accomplish this requirement, these works extend Transformer models input using new information obtained by external models. These information include: caption of objects detected in the image, sentiments extracted from the text and/or from the image, race and gender obtained from the images, among others. It is worthy noting that all results reported in this section were obtained in the HMDC dataset.

The most common external information used is object/image captioning, as is done in [81]. The architecture proposed in this work employs three pieces of information from each meme: 1) the text–extracted using an OCR; 2) a list of objects and their labels–extracted from the image using the Faster R-CNN model [82]; and 3) a caption automatically generated to the image. This last information is obtained using a CNN and a Decoder Sequencer trained using the MSCOCO dataset [83], which contains 123,000 images and five reference captions per image. These three sources of information are the input to a triple relationship network, which is a Transformer network created to model cross-modality relationships between image features and the two textual features. The authors employ both single-stream (SS) and dual-stream (DS) visual-linguistic Transformers as base models. In the former, the two modalities share the same input bus, leading the Transformer layers to operate on a concatenation of inputs. In the latter, image and text are not concatenated at the input level so as each modality has its own input bus.

The VisualBERT [84] is the SS model employed, while ViLBERT [85] is the DS one. When comparing the results reached by both models to other monomodal and multimodal approaches, the authors point out VisualBert using the three sources of information as the most performing model: 73.98 AUCROC in the test set. It is important to mention that VisualBert employing image, text and caption outperformed its version that takes into account only text and image. These results confirm the hypothesis that augmenting the model with external information related to the meme improves reasoning and understanding needed to solve the challenging hateful meme detection problem.

The work described in [8] focuses on adding tags of objects detected in the image to the input of Transformers, besides the textual and visual features. Moreover, different from the previous work, they create ensemble of classifiers instead of working

with only individual models. They adopted three multimodal models based on the Transformer architecture: LXMERT [86] (DS), UNITER [87] (SS) and OSCAR [88] (SS). The first is a model that processes images and text independently using unimodal encoders. The combination of the unimodal representations occurs through a cross-attention module. The second uses a self-attention mechanism to combine text and image input, creating a common space. Finally, the third works with a triple entry, formed by a sequence of words, tag of objects detected in the image, and features of the object regions. In their initial experiments UNITER attained better results. As a consequence, they created ensembles of models using only variants of UNITER. One of the variants adds to the input of the model classes predicted by the YOLO9000 image object detector [89]. The most performing ensemble combination was obtained by grouping a set of three different ensemble models. This approach attained 80.53 AUCROC in the test set, improving the results presented in [81]. According to the authors, the use of an ensemble of models is the key issue for this approach achieving high AUCROC rates.

An ensemble of UNITER variants is also the best performing solution proposed in [12]. They add caption information inferred from the meme image by a model trained on a different corpus. The original meme text is first extracted using OCR. Then, the Show and Tell model [90] is used to generate a new image caption, which replaces the original text. This way each image will be duplicated, one associated with the original meme text and the other associated with the text generated by the Show and Tell model. The ensemble of UNITER models is combined in a late fusion approach performed by averaging the models probability when assigning labels to the instances. This work focuses on increasing model diversity and data augmentation as approaches to try to improve the results on detecting hateful memes. The strategy used to generate new text for the meme image provided a significant effect, doubling the size and diversity of the data set. However, the AUCROC achieved was 79.43, inferior to the results shown in the previously described work. It is important to assess how closely the new text aligns with the original meaning of the meme as a way to increase the results.

In [11], ensemble of models is also employed. The meme text is extracted using OCR, while Detectron2 [91] is used to detect and extract objects from the meme

image. Different models are used for the extraction of visual features. Then, visual features, the object tags predicted from the regions of interest of the image (ROIs), and the meme text are fed into five different models focused on creating an ensemble of classifiers. The models employed are: a) ERNIE-ViL (DS) (both small and large) [92], based on ERNIE and VilBERT; b) VisualBERT; c) UNITER [87]; and d) OSCAR [88]. Each model generates a multimodal representation combining image and text. In addition, 3-5 variants of each model were generated by diversifying the features used to train the models. Finally, an ensemble of 19 members was obtained, whose outputs are combined using a late fusion function, such as simple averaging, rank averaging, etc. The best result reported by the authors in the HMDC dataset is 81.56 (AUCROC) for the test set. This result is superior to the rates reached by the three previous methods.

Besides image/object captioning, the complementary information can be generated based on the meme text. For instance, the work [93] initially follows the common procedure observed in previous works. Precisely, image modality is represented by three sources: the whole image–ResNet-152 is used to extract these features; images of all objects contained in the main image–detected by Faster-RCNN; and the ROIs position embedding present in the main image. For the textual features, these are extracted by BERT. The complementary textual features are provided by the Spacy [94] tool as follows. Noun phrases are obtained from the text. Then, keywords are picked up by filtering out irrelevant word. The resulting text representation combines BERT and Spacy representations. The authors propose a Complementary Visual and Linguistic (CVL) network composed by ViLBERT and VisualBERT, both receiving image and text as input. CVL implements the early fusion approach by concatenating the two models output, which is then fed into a fully connected network responsible for predicting whether the meme is hateful or not. The AUCROC achieved by this approach was 78.48, lower than the rates reached by previous works that added external visual features.

In order to further exploit the multimodal understanding, the authors in [10] propose to extend the information provided to the models by adding text and image sentiment analysis. Their hypothesis is that sentiment analysis is a related task that may enhance the ability of the models to identify hateful memes. To

accomplish this objective, VisualBert is used to extract a multimodal representation of the meme text and image. Two additional models are generated: RoBERTa [95], which extracts sentiments from the text; and VGG [67] to extract sentiments from the image, both pre-trained with sentiment analysis datasets. The three obtained representations are concatenated following the early fusion approach, and the result of this concatenation is fed into a fully connected network to classify the meme. This method is compared to other approach proposed by the authors that combines the multimodal representation generated by VisualBERT [84] to a description text of all objects present in the image provided by a caption generator. The captions serve as input to BERT. The fusion and prediction processes follow the same strategy used with the sentiment analysis information. However, the best results were reached by the method that did not take into account sentiment analysis: 74.00 AUCROC, below previous work's. One possible reason for this result may be the fact that sentiment obtained from the images may not have a meaning by itself closely related to the meme message. In the case of text, the meme words may mostly not have the sentiment information clearly defined.

Finally, the work presented in [96] goes still further on extending information provided to Transformers besides the image and text description of memes. For this, Google Vision Web Entity Detection is used to generate a description of the image based on its context. Moreover, The FairFace [97] classifier detects and predicts race and gender of the majority of people whose head is not obscured in the image. These new tags feed an extended ERNIE-Vil [92] model, which creates a detailed representation of this entire set of information. After that, the combined representation is used as input to a fully connected network to assign a label to the meme. This work attained the highest AUCROC (84.50) among all papers discussed in this chapter. Therefore, the use of additional information related to the meme increases the power of transformer-based models, since using extended tags was the main characteristic of this work. Considering the improved results achieved, this seems to be a more effective way to tackle the intrinsic multimodality of hateful memes.

3.4 Discussion

The hateful meme detection task cannot be considered a trivial problem. The small number of works dealing with this topic and the usually weak performance attained confirm this fact.

Non-attention Mechanism-based Approaches		**Attention Mechanism-based Approaches**	
Restricted Methods		**Restricted Methods**	**Extended Methods**
Hand-crafted Feature Extraction-based Methods	Auto-feature Extraction-based Methods	Auto-feature Extraction-based Methods	Auto-feature Extraction-based Methods
(Fersine et al, 2019) [61]	(Sabat et al, 2019) [70], (Vlad et al, 2020) [73], (Suryawanshi et all,2020) [27]	(Velioglu and Rose, 2020) [78]	(Zhou et al, 2021) [81], (Sandulescu, 2020) [12], (Muennighoff, 2020) [11], (Zhang et al, 2020) [93], (Das et al, 2020) [10], (Zhu, 2020) [96], (Lippe et al, 2020) [8]

Table 1 – Reviewed works grouped according to the proposed taxonomy.

In this section, we discuss some choices made in current literature and evaluate other scenarios that can boost the results. Table 1 summarizes works reviewed in this chapter grouped according to the proposed taxonomy.

- **Datasets:** There are works that created their own datasets, precisely [61], [70], [27], and [73]. Some of these authors provided guidelines to the annotators, composed by examples and clear definitions, to help eliminate doubts during the annotation process. On the other hand, there are authors who did not define guidance to the annotators to avoid influencing their judgment. Finally, there are also cases whose annotations were performed by search engines. However, in all cases the authors reported problems related to the annotation process, since defining whether a meme is aggressive can be difficult, even for a human being. Therefore, creating a new dataset is a very challenging task. The Hateful Memes Dataset (HMDC) [24] helped to reduce this limitation. In Table 2, a summary of all datasets used by the reviewed works is shown.

- **Data Augmentation:** The publicly available datasets are often small-scale datasets. Due to this limitation, some studies focused on data augmentation, which was conducted especially in three ways:

  1. **Using object detector:** In this approach, an object detector extracts objects present in the meme image, usually providing the following infor-

!ht

Dataset	# of Instances	Classes	Works
[61]	800	sexist, noSexist, sexistIronic, sexistAggressive	[61]
[73]	1K6	hate, noHate	[73]
[70]	5020	hate, noHate	[70]
[27]	743	offensive, noOffensive	[27]
MSCOCO [90]	123K	5 captions per image	[81]
HMDC [24]	10K	hate, noHate	[8], [10], [93], [11], [96], [81], [78]
MMHS150K [98]	150K, only 16K used	hate, noHate	[12]
Memotion Dataset [79]	14K	notFunny, veryTwisted, hatefulOffensive, notMotivational	[78]

Table 2 – Datasets employed in the reviewed works. Several works use more than one dataset.

mation: a) Region of the image where the object was observed; b) object name; and c) position of the object in the image.

2. **Using caption generator:** In this case, a caption generation model creates a new text from the meme image, which is concatenated with the old text or is used as a superimposed text combined with the meme image to generate another meme.

3. **Using regions from meme image:** In this approach, regions are extracted from meme image to create new images, whose features are extracted to increase the number of image data.

These three approaches increase the number of input data. However, in the first and second cases, they are too dependent on the models' ability to interpret objects' presence and the image's content to generate the caption text. When taking into account the number and variety of memes, the two first approaches may not cope well with these characteristics. To minimize this risk, the use of a set of object detectors and different caption generators could in-

crease the chance of success. The third approach seems to be more promising, since it does not depend on object detection or caption generation.

- **Evolution of the architectures:** Although currently there is only a few studies focused on dealing with hateful meme detection, and most of them made public very recently, it is possible to observe the proposed solutions' evolution. We can define this evolution in four phases:

  1. **Focus on Feature Extraction:** In this phase, the works have evolved according to the evolution of the feature extractors of both image and text. Starting with classical extractors and then evolving to deep neural networks, whose results have shown their capacity to extract complex semantic information.
  2. **Focus on Data Augmentation:** Here, works focused on increasing the amount and diversity of data (image and text of the meme) are highlighted. Different strategies were applied, including object detectors, image caption generators, etc.
  3. **Focus on deepening the semantic meaning:** In this phase, the objective is to use better the multiple information generated in previous phases. This is accomplished by using the Transformer models, since these models present high capacity to extract semantic information more deeply.
  4. **Focus on Ensemble of Models and Extended Information:** The more recent works massively employ ensemble of models, combining different architectures of Transformers, fusing multiple results to finally carry out the task of predicting the meme class. These models also extend the input of Transformer-based models by adding information such as: gender, sentiment analysis, race, among others, to better deal with multimodality.

3.5 Final Remarks

Identifying memes with hate speech is a significant challenge in the real world. This chapter presents a review of the research focused on this task. The proposed

methods are explained in detail, with their results thoroughly analyzed. Similarities among the current approaches were identified, leading to the formation of distinct groups, which are represented in a proposed taxonomy. Describing each method, along with its strengths and areas for improvement, we provided insights into the progression of techniques and highlighted the most promising future directions.

The key takeaway from this review is that relying solely on the meme's image and text is insufficient for achieving high performance in hate speech detection. Notable improvements were observed when researchers incorporated additional information from the meme and used more sophisticated models, such as ensembles of Transformer-based architectures. However, these models often require substantial computational resources for training and incorporating additional information besides the textual and visual components of the memes, making them complex and challenging to reproduce. Thus, two promising strategies emerge: first, maximizing the extraction of direct information from the meme, and second, utilizing models capable of generating deep and comprehensive representations of all the extracted data.

The following chapter outlines our second main contribution, which presents a new method aimed at improving the performance of existing multimodal transformer models without requiring additional information. This is achieved by integrating a specialized module, known as Compact Parameter Blocks, into the encoders of the Transformer models.

4 Adding Compact Parameter Blocks to Multimodal Transformers to Detect Harmful Memes

The development of tools that can detect and eliminate harmful memes before they reach a wide audience and cause harm is a very important issue. It was discussed in the previous chapter that despite being multimodal data, early studies focused on identifying harmful memes using monomodal methods. For instance, clustering techniques [99], analyzing the textual content using methods such as bag of words [100], N-grams [101], similarity of topics [102], etc. The use of visual data in a monomodal mode was also explored. For instance, object identification [103], web entity detection [104], caption generation [105], direct extraction of visual features [106], among others. However, current studies are devoted to tackle the task using multimodal representations by taking into account the text and image components of memes together.

It is important to mention that off-the-shelf tools designed for general multimodal analysis might not be sufficient to unravel the inherent meaning of these memes due to several reasons [107]. Firstly, memes frequently rely on their context, which can greatly influence their interpretation. Moreover, their visual and textual components often lack a direct connection or correlation. The memes message can only be fully understood by analyzing both elements together. Modality, as stated in [6], refers to the manner in which an event occurs and the emotions it evokes. By combining various modalities, such as sounds, visuals, scents, and textures, we are able to gain a more comprehensive understanding. Together, different modalities enhance the meaning and significance of the information received.

In this context, the high performance rates reached by the transformer architecture in NLP tasks [7] motivated many researchers to also employ these models to detect harmful memes in a multimodal approach. These techniques use the transformer encoder component to gather both textual and image characteristics from

memes. Afterward, these merged features are used as the input to a classifier. The effectiveness of this strategy arises from the attention mechanism integrated into the encoder section of these models. This mechanism excels in grasping a comprehensive representation that arises from merging text and image elements. Due to the substantial computational demands and restricted data availability, most of these techniques employ transfer learning approaches when deploying these models. Despite yielding positive results, there is space for enhancing their performance.

The method presented in this chapter aims to enhance the efficiency of multimodal models that utilize the transformer architecture along with transfer learning techniques to detect harmful memes without requiring additional information. To achieve this goal, the proposed method involves integrating adjustable Compact Parameter Blocks (CPBs) into the encoder sections of these models.

The incorporation of CPBs improves performance by dynamically adjusting the weight distribution within the attention mechanisms of the encoder blocks. These blocks act as intermediate feature processors, capturing finer details while reducing reliance on excessive parameters. This is analogous to low-rank adaptations in transfer learning, where task-specific adjustments enhance efficiency without requiring full model retraining [108]. Additionally, CPBs enable the model to learn more targeted and robust representations, particularly for tasks involving subtle multimodal cues such as harmful meme detection.

While the inclusion of CPBs increases model complexity, this additional complexity functions as a form of implicit regularization. By redistributing learning capacity across compact parameter spaces, CPBs mitigate overfitting by preventing the model from overly relying on specific connections or neurons. This behavior mirrors regularization techniques such as dropout or weight decay, striking a balance between model complexity and generalization.

Theoretically, CPBs enhance the model's latent space representation by refining the features extracted by attention mechanisms. This process is similar to the structured latent representations produced by variational autoencoders (VAEs), which learn meaningful variations in data [33]. By prioritizing essential features and minimizing noise, CPBs improve the quality of the latent space, making the

model more effective in capturing the multimodal nuances required for harmful content detection.

The operation of CPBs is further inspired by autoencoders, as described in [109]. Autoencoders are neural networks primarily employed in unsupervised learning to convert input data into a compressed representation, often referred to as the "latent space." This transformation involves a series of layers that gradually extract abstract features, generating a condensed but informative representation that retains essential data characteristics while discarding nonessential details. However, unlike traditional autoencoders, CPBs achieve compression and feature refinement without modifying the existing loss function. Instead, CPBs integrate seamlessly into the encoder-decoder architecture, leveraging the same supervision mechanisms (e.g., cross-entropy or binary classification loss). This ensures that compression and reconstruction processes enhance intermediate feature processing without altering the model's original optimization objectives.

By incorporating CPBs, the method aims to shift some of the harmful meme detection responsibilities to the attention mechanisms embedded within each encoder block. This adjustment is expected to improve the model's accuracy, efficiency, and generalization capabilities in detecting harmful content.

In this chapter, we investigated four distinct datasets, each using predefined criteria to classify memes as harmful or non-harmful. Similarly, we employed two pretrained models for meme classification. Considering the limitations of these models, including inherent inaccuracies stemming from pre-training data, architecture, and feature extraction methods, our proposed approach seeks to alleviate these limitations. We hypothesize that the introduction of CPBs will enhance model performance and reduce classification errors.

In the rest of the chapter, Section 4.1 reviews the related research on detecting harmful memes. In Section 4.3, the proposed approach is outlined. Section 4.4 details the experimental results, demonstrating the advantages of our method. Section 4.5 presents and discusses our findings. Section 4.6 addresses the method limitations. Finally, Section 4.7 summarizes the conclusions.

4.1 Related Work

This section will cover the latest methods designed to identify harmful memes. Since several related works were previously discussed in Chapter 3, only works not discussed in the previous chapter are described here.

The approach proposed in [13] focuses on improving the interaction among features extracted from the text and the image of the meme. They developed a method called Hate-CLIPper to identify hateful memes through multiple forms of media. This method utilizes the encoders of Contrastive Language-Image Pretraining (CLIP) [51]—a visual-linguistic model—to obtain aligned image and text representations, which are then fused through bilinear pooling (outer product) to create a Feature Interaction Matrix (FIM). This FIM representation models the correlations between the dimensions of the image and text feature spaces, allowing a simple classifier to achieve high performance on hateful meme classification using only the FIM representation. Hate-CLIPper does not use additional information as input, however, the FIM representation incorporated various details for each meme, such as indications of whether it contained attacks directed at individuals with disabilities, nationality, race, religion, or gender. Additionally, a descriptive text for the image of each meme was generated, and all this data was utilized in the model training process. The design of the architecture involves a component called the CLIPMLP module. This module includes projection layers that serve the purpose of synchronizing the representations of text and images. The proposed architecture achieved AUC-ROC of 85.80 on the FBHM dataset.

The CLIP model is also employed in [110] to detect harmful memes, as well as to identify their intended recipients. The authors validated their method using two datasets, namely Harm-C and Harm-P [111]. Their technique involves processing a meme using Google's OCR Vision API to extract embedded text. Then, CLIP encodes text and image pairs in order to capture the essence of the meme. Additionally, the system identifies faces and suggests objects, capturing attributes of subjects within the image. VGG-19 [112] is used to encode visually relevant regions, while DistilBERT [113] encodes textual elements. Given the abstract and contextual nature of harmful memes, the approach suggests that incorporating identified objects and at-

tributes enhances the understanding of high-level meme concepts, thus effectively capturing important contextual details. The subsequent stages involve merging image and text representations with CLIP features, utilizing intra-model and inter-modal attention. This results in a context-sensitive multimodal representation that predicts the level of harm in a meme and identifies potential targets. The authors achieved accuracy of 83.82 and a F1 score of 82.80 for the Harm-C dataset, and accuracy of 89.84 and a F1 score of 82.80 for the Harm-P dataset.

Each method discussed in this section, as well as most of the methods discussed in the previous chapter, shares a common feature: they all incorporate additional information such as new meme caption generation, gender and race analysis, and more, alongside the original meme content, which includes both text and images. In this thesis, however, our proposed method only relies on the objects that can be identified in the image and the text associated with the meme. In the following section, we introduce our proposed approach.

4.2 Methodology

Figure 17 illustrates a generic framework for detecting harmful memes using multimodal transformer models. The input consists of memes that typically combine both text and images. The framework extracts features from these memes in two parts: text representation and image representation. The text representation extractor processes the detected text, while an object detection module identifies objects within the meme image. The detected objects are then passed to an image representation extractor, which captures visual features.

The extracted text and image features are then combined to create an integrated multimodal representation of the meme. This fused representation is fed into a multimodal transformer model, such as VisualBERT [84] or VilBERT [85], which serves as the core feature extractor. Considering the proposed taxonomy shown in Figure 12, our approach is based on a transformers architecture using the attention mechanism. We adopt an expanded approach to the data, where features are automatically extracted. The key contribution of this work, CPB, is incorporated

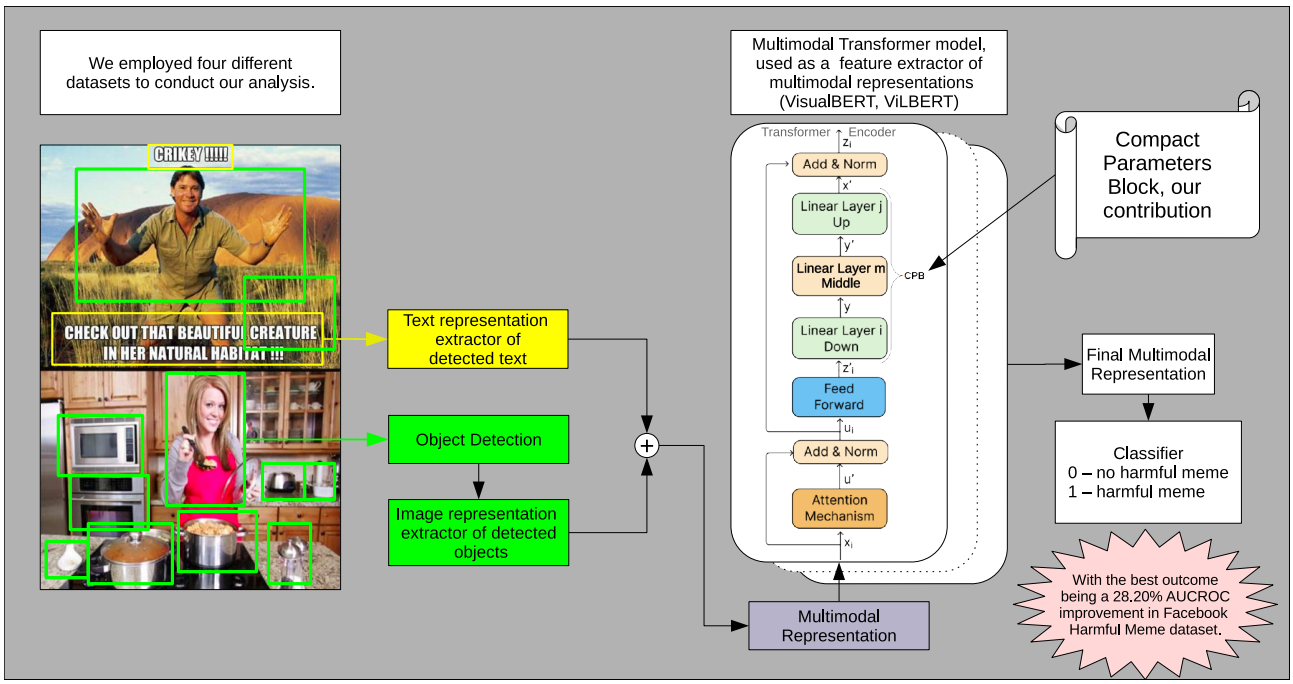

Figure 17 – General framework used for evaluate CPB approach. Source: Author.

into the encoder segments of the transformer to enhance its performance without requiring additional data. The model generates a final multimodal representation, which is used by a classifier to determine whether the meme is harmful (1) or non-harmful (0).

4.3 The Compact Parameter Blocks Approach

We work with two different architectures of multimodal transformer models. The first is called Single-Stream (SS) and it involves combining all the multimodal data at the model's input, separated only by tokens. Examples of models from this category are: ImageBERT [114], VisualBERT, ERNIE-Vil, Unicoder-VL, and VL-BERT. The second is called Double-Stream (DS), having separate inputs for each modality. LXMERT [86], VilBERT, and CLIP are some models in this category. Our goal is to demonstrate that the CPB method can be applied to any of these models, as the CPBs are added to the encoder components, which are a crucial component in both SS and DS methods. All examples are shown taking into account two modalities, since the memes investigated in this work are bimodal data. In addition, we add CPBs to the SS model VisualBERT, as well as to the DS model VilBERT. These

models were selected because they represent a diverse architectural paradigm in a multimodal context. This strategic choice enhances the generalizability of our findings to a broader spectrum of models with similar architectures.

The rigorous initial training with extensive datasets allows multimodal transformer-based models to undergo a comprehensive learning process. During this phase, they become familiar with complex procedures. When these pre-trained models are applied to solve tasks using small-sized datasets, the most common procedure is to use transfer learning, where most of the model remains unchanged, except for the classifier, which adapts to the new task.

Our proposed methodology involves incorporating an unfrozen block, termed CPB, with an autoencoder like structure [109, 115], into the transformer model encoder during the transfer learning phase. We hypothesize that this integration will substantially enhance the performance of transformer models, particularly in the multimodal task of establishing relationships between text and image components within memes to accurately classify their harmfulness. This modification is expected to yield several key benefits to the original encoder operations, such as:

- Enhanced Encoding Efficiency: The CPB, positioned at the encoder blocks, compresses the input text and image into a lower dimensional latent space representation. This compressed form encapsulates the most representative information, while filtering out noise and redundancy.

- Optimized Attention Mechanisms: By receiving a refined, information-dense representation of the input data, the transformer's attention mechanisms can more effectively focus on critical aspects, thereby improving contextual understanding and interrelationships between text and image elements.

- Regularization Effect: The inclusion of the CPB introduces a regularization effect, mitigating the risk of model over-fitting on training data and consequently enhancing its generalization capabilities to unseen data.

The integration of the CPBs into the encoder block of a transformer model presents a promising direction on achieving significant gains in efficiency, representation quality, and robustness. Figure 18b provides a detailed illustration of the CBP

architecture. In the following subsection, we will provide a detailed explanation of the implementation of CPB within the VisualBERT model.

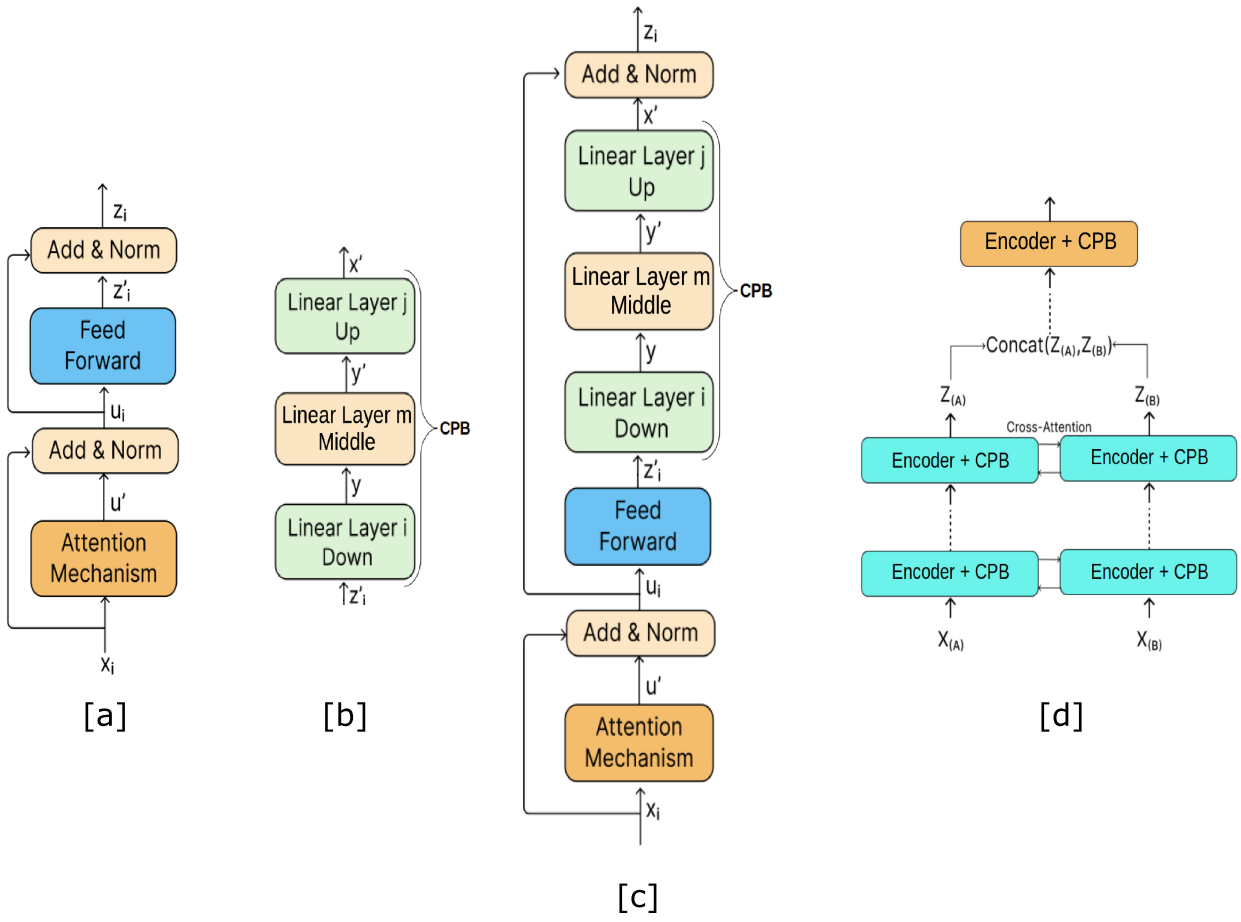

Figure 18 – (a) Visual representation of the encoder block of the VisualBERT model, focusing on the attention mechanism. It is at this block that we will introduce the structure of the CPB. (b) The CPB structure is a set of layered components arranged in a sequence where they first decrease and then increase input data. (c) A VisualBERT encoder featuring a CPB integrated within its self-attention mechanism. The CPB is intended to extract the output from the current encoder block and generate a better representation, which will subsequently serve as input for the attention mechanism in the following encoder block. (d) VilBERT model encoder with a CPB. Here, $X_{(A)}$ represents the textual information, while $X_{(B)}$ represents the visual (image) information. These two sets of data are combined and fed as input to the final encoder blocks, in which we also add CPB blocks within their architectural design. Source: Author.

4.3.1 CPB in Single-Stream (SS) multimodal transformer model

In order to present the CPB approach, we use VisualBERT as a foundation. This is a SS model pre-trained with the MSCOCO [116] dataset. Its architecture resembles

that of BERT, but it uses a more complex input sequence. This input sequence is comprised of three main components. Firstly, it includes representations of objects found within the image, which are detected using object detection techniques like Faster R-CNN [117]. Secondly, there are tokens that mark the beginning and the end of these representations. Lastly, the textual content is incorporated. Moreover, the model generates positional representations for each element in the input sequence. This input is divided into segments, with items enclosed by special symbols considered part of the same segment. The segment information is also integrated into the model's input.

Considering that the CBPs are added to the encoder module of VisualBERT, in Section 2.3 we describe its encoder blocks. As previously mentioned, the attention mechanism is a crucial component of the encoder blocks. Therefore, understanding how this mechanism operates is especially important. For more information related to the attention mechanism, refer to the reference provided in [7]. Since in Section 2.3 we describe a summary of this module, in this point we only introduce the CPB in the encoder context.

A CPB is composed of three different linear layers: 1) Down; 2) Middle; and 3) Up, as depicted in Figure 18b. The objective is first decrease and then increase the input data. These layers collaborate in a step-by-step manner to condense and then expand the input embeddings. This entire process helps to create a reliably stable and uniform representation.

The addition of a CPB occurs as follows. The input representations z'_i produced by the encoder block before the residual connection (Equation 2.8) are passed through a down linear layer, i, creating the output y, as shown in Equation 4.1. In this equation, W_i is the randomly generated weight matrix with dimension $(d_{z'}, d_y)$, with $b = 0$ and $(d_y \approx \frac{d_{z'}}{10})$. Therefore, it takes the input data z'_i, and maps it into a lower-dimensional representation. The objective of this layer is to reduce the input dimension, similar to the approach used in autoencoders [109], focused on several advantages this approach present. First, lower-dimensional embeddings tend to retain the most important aspects and connections within the data while filtering out less important details. As a result, these representations are simpler to grasp, making it easier to draw meaningful insights from the data. Additionally, it aids

in reducing the impact of noise. By compressing the data into a more condensed form, the influence of noise and less significant features is minimized. This, in turn, results in more robust and reliable outputs.

$$y = z_i' W_i^T + b \tag{4.1}$$

The next step of the CPB is the middle layer, as shown in Equation 4.2, located in the core of the CPB. This layer has fewer neurons (dimensions) than the input data. Consequently, it is expected to provide a compressed or latent representation of the input data, forcing the CPB to capture the most essential features of the data.

$$y' = y W_m^T + b \tag{4.2}$$

In this equation, y' feeds a new linear layer m, where W_m is the randomly generated weight matrix whose dimension is $(d_{y'}, d_y)$, with $b = 0$ and $(d_{y'} = d_y)$.

Then, in the last CPB layer, y' feeds a new linear layer j. The objective here is to use the compressed representation to attempt to reconstruct the original input data from this lower-dimensional representation. Therefore, the goal is to generate an output that is as close as possible to the input. Equation 4.3 shows the operations performed in this layer.

$$x' = y' W_j^T + b \tag{4.3}$$

In this equation, x' is the CPB output, W_j is the randomly generated weight matrix whose dimension is $(d_{y'}, d_{x'})$, with $b = 0$ and $(d_{x'} \approx 10 d_{y'})$. Finally, the second encoder normal layer receives x' and the residual u_i, as shown in Equation 4.4.

$$z_i = LayerNorm(u_i + x') \tag{4.4}$$

Figure 18c provides a visual representation of how a CPB is integrated into a VisualBERT encoder. The process continues with the next encoder block, an so on. It is important to mention that in terms of SS models, which use a single input stream per modality, there are few choices to integrate the CPB. However, the DS models have a broader range of options, as discussed in the next section.

4.3.2 CPB in Double-Stream (DS) multimodal transformers model

When working with DS models, where the input streams of modalities are separated, there are three potential options for incorporating CPBs. These options include applying CPB to the text stream only, to the image stream only, or to both streams simultaneously. In our experiments using VilBERT (pre-trained on the Conceptual Captions [80] dataset), the best results were achieved when CPBs were applied to all three options. Therefore, in this work we added CPBs in three parts simultaneously, as it is illustrated in Figure 18d.

In the next section we demonstrate through experiments that the CPB approach yields positive results. To achieve this, the detection of harmful memes will be tested using both VisualBERT and VilBERT.

4.4 Experiments

In this section, we describe the experiments conducted in this work and the results attained. The experiments are focused on examining the impact of incorporating CPBs into the models VisualBERT and VilBERT. First, however, we discuss the datasets investigated.

4.4.1 Datasets

We employed four different datasets to conduct our analysis. These datasets all share a common theme: they contain instances of memes that feature images overlaid with text. Although some of these datasets contain information for multiple categories, our approach focused solely on binary classifications for all of them. Specifically, we assigned two possible labels to each meme instance: 0 to indicate non-harmful memes and 1 to the harmful ones. We made this choice primarily because it aligns with the common definitions used in baseline models. The datasets are described as follows.

[a] [b] [c] [d]

Figure 19 – (a) An instance of a harmful meme extracted from the FBHM, (b) MMHS150K, (c) MultiOFF and (d) the MEME datasets respectively. Source: Author

4.4.1.1 FBHM [24]

This dataset is the same called Hateful Memes Dataset (HMDC) in our previous chapter. It includes a total of 10,000 memes, including memes with content representing different types of aggression aimed at groups that are legally protected. One common form of aggression represented is dehumanization, where individuals are compared to non-human objects or animals. The dataset is composed of memes with various forms of dehumanization. In addition, harmful memes are categorized into specific groups, including an "other" category. However, in all works that investigate this dataset, each meme is assigned to only one category of attack, composing a binary classification problem. One example of a harmful meme from the FBHM dataset can be seen in Figure 19a.

4.4.1.2 MMHS150K [98]

This dataset is composed of 149,823 memes, which are divided into two categories: memes that contain hate speech and those with no hate speech. Like the previous dataset, it offers files containing the written content of every meme. One example of a meme from the MMHS150K dataset is shown in Figure 19b.

4.4.1.3 MultiOFF [27]

It includes 743 memes centered around the 2016 United States presidential election. These memes have been categorized as either offensive or non-offensive. An example of a meme from this dataset is shown in Figure 19c.

4.4.1.4 MEME [61]

It is a benchmark dataset composed of 800 memes with sexist and non sexist content, labeled considering the visual and/or textual aspects. This dataset also includes both images and related texts. One example of an instance from the MEME dataset is shown in Figure 19d.

4.4.2 Experimental Protocol

In our experiments, we used all four previously mentioned datasets. Table 3 summarizes their data distribution and partition. While it is worth noting that not all of these datasets are balanced, we followed the baseline approaches by not undertaking any measure to address this imbalance.

Table 3 – Summary of the datasets used in the experiments in terms of their division into training, validation, and testing sets, along with the counts of samples in each class.

	FBHM	MMHS150K	MultiOFF	MEME
Train	8500	135000	445	600
Val	500	5000	149	100
Test	1000	9823	149	100
Total	**10000**	**149823**	**743**	**800**
Hate	3756	36978	440	369
Not-hate	6244	112845	303	431

In terms of the two multimodal models used in our experiments, for VisualBERT we used the model with its 12 encoder blocks in conjunction with a single dense layer that handled meme classification. A CPB was added in each encoder block. In its turn, VilBERT is composed of 12 encoder blocks for text, 6 encoder blocks for images, and 6 encoder blocks that receive the output embeddings from both

text and image encoder blocks. Again, in each of these encoder blocks we have incorporated a CPB.

To train VisualBERT and VilBERT on various datasets, we conducted training for a total of 22,000 steps. The number of epochs required for this process varied depending on the specific dataset used. As observed in Figure 18b, CPBs are composed of linear layers. The number of parameters P in a linear layer is determined according to the following equation:

$$P = (inSize + 1) * outSize \tag{4.5}$$

In this equation, $inSize$ represents the number of input features or neurons in the previous layer, while $outSize$ represents the number of neurons or units in the current linear layer.

Table 4 – Details of CPB parameters distribution in the VisualBERT model.

	VisualBERT	Parameters
Layer i (in-out)	768 - 76	58444
Layer m (in-out)	76 - 76	5852
Layer j (in-out)	76 - 768	59136
Total by CPB		123432
Total Model	all 12 encoder blocks	1481184

Table 5 – Details of CPB parameters distribution in the Vilbert model.

	VilBERT	Parameters
Text Encoder blocks (12)		
Layer i (in-out)	768 - 76	58444
Layer m (in-out)	76 - 76	5852
Layer j (in-out)	76 - 768	59136
Total by CPB		123432
Image Encoder blocks (6)		
Layer i (in-out)	1024 - 102	104550
Layer m (in-out)	102 - 102	10404
Layer j (in-out)	102 - 1024	105472
Total by CPB		220426
Fusion Encoder blocks (6)		
Layer i (in-out)	768 - 76	58444
Layer m (in-out)	76 - 76	5852
Layer j (in-out)	76 - 768	59136
Total by CPB		123432
Total Model	all 24 encoder blocks	3544332

Table 4 and Table 5 summarize information related to the additional parameters resulting from the inclusion of CPBs in the VisualBERT and VilBERT models respectively. To determine the total increase in the number of parameters, we multiply the number of parameters in one CPB by the total number of encoder blocks in the model. This accounts for the additional parameters introduced by adding these blocks multiple times.

In terms of VisualBERT, its original model comprises 112,044,290 parameters. However, with the incorporation of CPBs, the overall number of parameters increased to 113,525,474. In essence, the incorporation of CPBs in VisualBERT resulted in a 1.30% expansion in the overall number of parameters compared to the original model. For VilBERT, its model initially is composed of 247,780,354 parameters, which increased to 251,324,686 due to the addition of CPBs. Therefore, there was 1.41% of expansion in the overall parameter number compared to the original model.

Our experiments were conducted using a research tool known as "Modular Multimodal Framework" (MMF).

4.5 Results and Discussion

In this subsection, we evaluate the impact in the performance of VisualBERT and VilBERT when CPBs are added to their encoder blocks. Moreover, we compare our results to the results attained by other methods available in the literature focused on identifying harmful memes. To determine the baselines, we observed the overview provided in [19] and [118]. Based on these references, we selected the works that achieved the top results for each dataset used in our experiments. The evaluation criteria are accuracy (ACC), F1 score, and AUC-ROC. It should be noted that some of the baseline studies may not provide all these metrics, but we will present the available metrics for each of them. Moreover, since all datasets used present imbalanced class distribution, we focus our comparison on the F1 Score results, due to the fact that F1 Score is a better metric when there are imbalanced classes in the evaluated problem. We justify the choice of baselines by

verifying whether they explore textual and visual features extracted from memes. This evaluation ensures that the selected models effectively capture multimodal information, allowing for a more comprehensive comparison of their performance.

Table 6 – Summary of the results obtained in the experiments conducted using the MultiOFF dataset.

MultiOFF Dataset			
Approach	ACC	AUCROC	F1 Score
Early fusion: Stacked LSTM/BiLSTM/ CNN-Text + VGG16 [27]	-	-	0.50
BERT, Faster-RCNN, Disentangled representations) [119]	-	-	0.65
VisualBERT	0.60	0.56	0.70
VisualBERT-CPB	**0.68**	**0.67**	**0.77**
VilBERT	0.61	0.62	0.71
VilBERT-CPB	**0.66**	**0.65**	**0.74**

We first analyze the results attained when investigating the proposed method, as well as the baselines, using the MultiOFF dataset. Table 6 summarizes the results obtained. In this specific dataset, when we evaluate performance based on the F1 Score, our approach surpasses previous studies, demonstrating superior results. Notably, the VisualBERT-CPB Model achieved the most substantial improvement, increasing 15.60% in performance when CPB was applied. It is also evident from these results that employing either the VisualBERT or VilBERT base models alone, without CPB, they still yield better results compared to the two baseline methods. In addition, the results also highlight the positive impact of incorporating CPB into these models.

In the context of this specific dataset focusing on American presidential candidates, VisualBERT outperformed VilBERT by a small margin. This performance difference can be attributed to the dataset's characteristics, which predominantly contains memes that feature both images of the candidates and their names in the textual component. This strong association between visual content and textual information aligns well with VisualBERT's single-stream architecture, where it combines both modalities in a unified manner. This synergy between image and text likely played a significant role in VisualBERT's better performance on this

dataset.

Noteworthy is the fact that this dataset represents a very challenging classification problem. This aspect is confirmed when we observe the low performance attained in general. The highest F1 Score was 0.77, reached by VisualBERT-CPB. The two baselines reached much lower F1 Score results. The work described in [27], which is not based on transformers, attained the performance of a method that randomly guesses the instance label (F1 Score = 0.50). The second baseline [119], reached results slightly higher than random guess (F1 Score = 0.65). The reason for this behavior may be the small size of this dataset: it contains only 743 instances. This is a serious limitation, preventing learning models to provide better results.

Table 7 – Results obtained in the MEME dataset.

MEME Dataset			
Approach	ACC	AUCROC	F1 Score
Late fusion - Several Hand-Crafted visual and textual features [61]	-	-	0.76
VisualBERT	0.86	0.92	0.85
VisualBERT-CPB	0.95	0.95	0.88
VilBERT	0.86	0.91	0.83
VilBERT-CPB	0.91	0.95	**0.90**

Despite containing only few more instances than the previous dataset, the results attained in the MEME dataset are much better in general. Table 7 shows these results. Among the investigated methods, our approach again surpasses the baseline studies, providing superior results. Here, the VilBERT-CPB Model achieved the highest improvement, increasing in 15.55% its F1 Score when CPB was applied. However, the F1 Score reached by VilBERT-CPB was only slightly superior to that from VisualBERT-CPB. One possible reason for this behavior is observed in [61]. These authors explain that in this particular dataset, the individual modalities possess substantial discriminatory power to allow distinguishing between hateful and non-hateful memes. Therefore, since the VilBERT model initially deals with modalities independently before integrating them, this characteristic may pose a slight advantage over the VisualBERT model, which combines both modalities into a unified representation as input since the beginning of the whole learning process.

On the other hand, it is also possible to observe from this table that employing either the VisualBERT or VilBERT base models alone, without CPB, still yields better results compared to the other methods. These results also highlight the positive impact of incorporating CPB into both models since VilBERT-CPB and VisualBERT-CPB outperform its original versions.

Table 8 – Results obtained using the MMHS150K dataset.

MMHS150K Dataset			
Approach	ACC	AUCROC	F1 Score
FCM (Feature concatenation model), Inception-V3, LSTM [98]	0.68	0.73	0.70
VisualBERT	0.74	0.62	0.68
VisualBERT-CPB	**0.76**	**0.77**	**0.75**
VilBERT	0.57	0.65	0.62
VilBERT-CPB	0.65	0.74	0.72

The VilBERT-CPB Model also achieved the highest F1 Score improvement (6.66%) when CPB was applied in the MMHS150K dataset, as shown in Table 8. It also increased 10.52% in ACC and 5,20% in AUC-ROC. However, VisualBERT-CPB reached the best performance. Again, our approach outperformed the baselines. Surprisingly, however, are the low classification rates attained in this dataset, since it is the largest one among all investigated in this chapter: it contains an extensive collection of memes, over 149,000. One possible reason for these results is the high class distribution imbalance: hate speech memes constitute only 24.56% of the dataset, making this dataset significantly more unbalanced than the others. In addition, the instances of this dataset were labeled by three annotators. According to [98], due to the subjective nature of the task and discrepancies among annotators, achieving high classification rates in this dataset is challenging.

Despite the challenges, VisualBERT-CPB achieved 75% of F1 Score, establishing a robust correlation between text and image for the classification of the memes. These results suggest that VisualBERT outperformed the VilBERT model, because it integrates text and image into a unified representation. However, the results also highlight the positive impact of incorporating CPB into both models.

Finally, the results obtained using the FBHM dataset are presented in Table 9.

Table 9 – Results obtained in the FBHM dataset.

FBHM Dataset			
Approach	ACC	AUCROC	F1 Score
Cross-modal Interaction of CLIP Features [13]	0.83	0.85	-
VisualBERT	0.64	0.56	0.65
VisualBERT-CPB	0.68	0.78	0.67
VilBERT	0.63	0.60	0.56
VilBERT-CPB	0.70	0.77	**0.71**

In this dataset, VilBERT-CPB also achieved the highest F1 Score increase (21,12%) when CPB was applied. It also outperformed the F1 score obtained by VisualBERT-CPB. Moreover, the results reinforce the positive impact of incorporating CPB into these models. However, neither of the two models was better than the baseline [13], which is the current benchmark for the FBMH dataset. In this baseline, the authors adopt a cutting-edge approach involving multimodal pre-training to establish a connection between images and text by representing them in a shared feature space. They reinforce the significance of modeling interactions between image and text features through an intermediate fusion process.

It is important to observe that, unlike our method, this baseline incorporates additional information, such as indications of whether the meme contains attacks directed at individuals with disabilities, nationality, race, religion, gender, etc. In addition, the FBHM dataset provides text and vision confounders. These confounders lead to changing a hateful meme to a not-hateful one, or vice-versa, by swapping either image or text only. This characteristic makes the dataset somewhat divergent from the typical meme content [120] and with a very challenging nature. Therefore, a model must be able to tackle multimodal reasoning to classify the meme and its confounder correctly.

Considering all the results provided in this section, it is possible to observe that incorporating the CPBs into VisualBERT and VilBERT models yields improvements to the original models. However, the addition of CPBs is apparently more beneficial to VilBERT, since this model reached the highest F1 Score improvement in three cases out of four by adding the CPBs. On the other hand, the success of adding these blocks seems to be influenced by the alignment between image and

text. This effect is evident in scenarios like the MEME dataset, where the focus is narrow: memes comprising images and text predominantly associated with attacks to women. In this context, the maximum AUCROC gain, considering both models, is limited to 4.2%. This suggests that the original model possesses sufficient complexity to address the problem, rendering the introduction of CPBs less beneficial.

The FBHM dataset presents instances with significant misalignment between text and image, due to the fact that this dataset has memes with identical images but different texts, and vice-versa. In this challenging scenario, the introduction of CPBs yielded the most substantial result: VisualBERT-CPB with 28.20% of AUCROC gain. This suggests that CPBs are more effective in capturing subtle multimodal relationships between meme images and text than the original models.

4.5.1 Significance Test

Considering the fact that the best results attained by our method were obtained in the FBHM dataset, we conducted a comparison of these results using the t-test [121], a statistical tool employed to measure whether the differences between two groups are statistically significant. There are various types of t-tests, such one-sample, independent, and paired t-test, each tailored for specific scenarios. In this work, we performed the paired t-test to compare the two investigated models, VisualBERT and ViLBERT, to their version with CPB enhancement. The comparison was conducted using three metrics: Accuracy, Area Under the ROC Curve (AUCROC), and F1 Score. The reported values represent averages over multiple experimental runs, assuming a normal distribution of differences.

The results summarized in Table 9 were used to determine if the inclusion of CPB significantly improved the models' performances. The null hypothesis (H0) posits no significant difference, while the alternative hypothesis (H1) suggests a significant difference. With a significance level of 5% ($\alpha = 0.05$), the p-values calculated for VisualBERT and ViLBERT are approximately 0.000000 and 0.008163, respectively. These p-values represent the probability of observing the obtained

results if the null hypothesis were true. Therefore, the p-values obtained for both models fall below the significance level, thus providing evidence to reject the null hypothesis. This indicates that incorporating CPB leads to statistically significant improvements in the performance metrics for both VisualBERT and ViLBERT.

The subsequent section will discuss the limitations of our methodology.

4.6 Method Limitations

While our proposed method demonstrates promising results, certain limitations must be acknowledged. Firstly, its applicability is confined to models utilizing transformer architectures, with a particular focus on the encoder block. Secondly, the incorporation of CPBs inevitably increases the number of model's parameters, as evidenced in Tables 4 and 5. However, it is worth noting that for the models employed in our experiments, this increase remained below 2% of the original number of parameters. Furthermore, the most significant performance gains were observed under challenging conditions, where the original model struggled to establish meaningful relationships between image and text components of the meme. On the other hand, the contribution of CPBs was less pronounced when the original model performed well. Consequently, a thorough cost-benefit analysis is indicated before integrating CPBs, as the potential performance improvement may not always justify the increased complexity.

Moreover, despite the significant progress made, ethical considerations persist. As noted in [122], human review of model outputs is recommended prior to full deployment. To foster trust and accountability within the moderation process, both automated and human decisions should be transparent and explainable, allowing users to understand and potentially challenge content moderation actions.

4.7 Final Remarks

In this chapter we proposed a novel approach to enhance the performance of multimodal models based on transformer architectures for detecting harmful content

in memes. Our contribution lies in introducing the CPB structure within the encoder block, a process that refines input data by compression and decompression, thereby improving the attention mechanism and introducing a data regularization effect.

To validate our hypothesis, CPB was integrated into two models, VisualBERT and VilBERT, representative of diverse multimodal architectures. Rigorous evaluation across four meme datasets demonstrated a consistent improvement in model performance with CPB integration.

In summary, this chapter concluded with a detailed evaluation of the impact of CPB enhancement on the performance of VisualBERT and ViLBERT models, using the four different datasets. A paired t-test was employed to statistically compare the models' performances across Accuracy, AUCROC, and F1 Score metrics. The results demonstrated that the inclusion of CPB significantly improved the models' performance, with p-values well below the significance threshold of 0.05 for both models. These findings provide strong evidence to reject the null hypothesis, highlighting the effectiveness of CPB in enhancing the models' capabilities.

While acknowledging the limitation of this approach's applicability solely to transformer-based models and the increased model complexity it introduces, we emphasize the significant potential for further exploration and refinement.

Finally, in terms of ethical considerations inherent to the content moderation process, we reinforce the use of a hybrid approach that combines automated tools with human oversight to maintain trust and transparency. This balanced strategy may help to keep ethical standards while allowing technological advancements to effectively combat harmful content in the digital context.

The following chapter outlines our third main contribution: it tackles the challenge of detecting aggressive memes by leveraging MLLMs for multimodal analysis. Using prompt engineering, it categorizes memes by complexity and evaluates model performance in five datasets.

5 Exploring the Performance of Multimodal Large Language Models in Detecting Aggressive Content in Memes

In previous chapters, it was mentioned that, despite the fact that the meaning of memes is highly dependent on their context, their images and text often have no clear connection. Therefore, a thorough understanding of both modalities is necessary to truly understand the message of the meme. This is crucial to provide accurate tools that do not attack freedom of expression, which is a legitimate and inviolable right.

In the context of multimodal analysis, MLLMs have become a major innovation due to advances in artificial intelligence and deep learning [123]. Models such as GPT-4V [14], LLaVA [18] and Gemini [17] are able to understand complex details that were difficult for older systems. Considering the improved ability of this models to work with text and visual data, we investigate in this chapter the possibility of using them to detect offensive content in memes, since it appears to be very promising. It is expected that MLLMs analyze both text and images in detail, identifying patterns and hidden messages that can be offensive. This can help create better tools for content moderation, making online spaces safer and more inclusive.

However, it is well known in the literature that the main way to use MLLMs in new tasks is by performing Prompt Engineering [124, 125], since traditional transfer learning techniques are often not possible. This involves creating specific instructions or questions for the models, helping them to give better answers or perform tasks more accurately, providing relevant and clearer answers. Therefore, in this work, we employ Prompt Engineering to investigate the following three Generative Models in the task of aggressive meme detection: GPT-4V, LLaVA and Gemini.

It is important to mention that the generalization capacity of machine learning models is often measured by evaluating how well they perform specific tasks.

However, this evaluation is difficult with MLLMs. The lengthy pre-training that these models undergo and the lack of clarity about the training data used raise concerns about data overlap — known as data leakage. This makes it difficult to know whether good results achieved on some datasets are due to real learning or due to prior data exposure. In order to try to deal with this problem, in this work we do not use data from the training split of the datasets investigated, only their test splits. This is a small effort to reduce the data leakage problem.

Finally, in order to broadly evaluate whether MLLMs are capable of truly tackle multimodal reasoning to perceive the inherent meaning of aggressive memes, we propose grouping aggressive memes into three multimodality reasoning levels. These groups, which will be detailed in Section 5.2.2, are intended to establish a scale that measures how easily aggressive content in memes can be recognized by individuals. This analysis takes into account the multimodal nature of memes. In summary, the first group involves memes whose both text and image are overtly aggressive and are expected to be immediately recognized as such. The second groups memes that have only one modality with aggressive content, requiring more attentive analysis, but that are still quickly identified as aggressive. The third group is composed by the most challenging cases, those where neither modality is independently aggressive, but their combination conveys aggressiveness, requiring more in-depth contextual analysis for an accurate identification. It is important to highlight that this categorization was performed by manual annotation. Memes from five different datasets were submitted to 8 annotators, who were instructed to categorize them into one of the specific groups of multimodality reasoning level.

In this chapter we seek to answer the following questions:

- RQ1: How can we effectively measure the performance of MLLMs in detecting aggressive content in memes across different levels of multimodality reasoning?
- RQ2: How can prompts be improved, along with meme itself, to help MLLMs to better identify aggressive content in memes?

The rest of this chapter is organized as follows: Section (5.1) describes related

work on aggressive meme detection, which were not reported in previous chapters. Section (5.2) explains our approach. Section (5.3) shows our experimental results. Section (5.4) discusses our findings. Finally, Section (5.6) presents conclusions.

5.1 Related work

The development of methodologies to detect aggressive content in memes is a growing field of research, specially due to new meme datasets [126, 127, 29], and to projects such as the Facebook's Hateful Memes Challenge [24] that have encouraged the creation of new ways to identify aggressive content in memes, especially hate speech. In this section, we detail some recent approaches whose focus is on exploring all the potential of multimodality in memes. It is important to mention that only works not discussed in previous chapters are described here.

In [128], the authors worked with prompts to try to improve the rates of hateful memes classification. Their goal was to investigate the performance of RoBerta by providing a monomodal prompt consisting of three texts: 1) the meme text; 2) a caption generated from the meme image; and 3) the text "This is [MASK]". During training, [MASK] takes the value "Good" or "Bad", and during inference, the model returns the percentage of each value. They used two datasets in their experiments, FBHM and HarM [129]. The results obtained were accuracy of 72.98 and AUC-ROC of 81.45 for FBHM, and accuracy of 90.96 with AUC-ROC of 84.47 for HarM. While the idea of this method is quite similar to ours, in our methodology we use a multimodal prompt (text and image) and Generative Models, unlike their use of RoBerta [130].

The automatic detection of misogynous content in meme is addressed in [131]. They advocate that it is crucial to explore methods for identifying this type of content from a multimodal perspective. The authors investigate four unimodal and three multimodal approaches to detect misogyny, using the dataset provided in[61]. The experimental results revealed that a combination of both text and visual elements is necessary for an effective classification. Additionally, a bias estimation and mitigation technique using Bayesian Optimization is proposed to correct bi-

ased model predictions by identifying specific elements in memes that could lead to unfair classifications. The proposed method shows improved accuracy, correctly classifying up to 61.43% of cases. Furthermore, the study highlights meme archetypes that pose significant challenges for existing misogyny detection systems and suggests that these should be further analyzed with diverse vision and language models to enhance future research in the field.

The work presented in [132] introduces an enhanced multimodal fusion framework which uses a congruent reinforced perceptron, inspired by human cognition, to better comprehend and reason about hidden meanings in memes. By dividing multimodal representations into primary semantics and auxiliary contexts and encoding them through a prefix uniform layer, the framework integrates these representations within a shared latent space. This approach strengthens the detection of subtle metaphors and implicit meanings behind hateful memes. Experimental results on the benchmark datasets Harm-C and Harm-P [110] achieved accuracy of 85.03 and F1-Score of 84.24 on the Harm-C dataset, while accuracy of 92.68 and F1-Score of 92.66 were obtained in the Harm-P dataset.

Finally, the authors in [133] introduce a novel multimodal dataset called MIMOSA (MultIMOdal aggreSsion dAtaset), which includes 4,848 memes designed to detect the targets of aggressive Bengali memes across five categories: Politics, Gender, Religion, Others, and Non-aggressive. A Multimodal Attentive Fusion (MAF) framework is proposed, utilizing CLIP for image feature extraction and a BERT model trained on Bengali for textual feature extraction. A distinguishing feature of this approach is the attention mechanism applied before the fusion of textual and visual data, where text features serve as the Query (Q) and image features as the Key (K) and Value (V) within the attention fusion mechanism. Experiments on the MIMOSA dataset demonstrated that MAF significantly outperformed eleven state-of-the-art unimodal and multimodal baselines, achieving accuracy of 0.741, weighted F1-Score of 0.642, and mean multimodal average error of 0.645. These results highlight the effectiveness of the MAF approach in leveraging multimodal context for identifying aggression targets within memes.

The methods discussed in this section have one thing in common: they try to correlate visual and textual information to obtain a thorough understanding of both

modals in order to truly understand the message of the meme. However, they do not differentiate aggressive memes in terms of levels of multimodality reasoning. In this chapter, this aspect is taken into account. In the following section, we will provide a comprehensive explanation of our methodology. Identifying aggressive memes is a complex and nuanced issue, and evaluating it within the context of Generative Models requires some adaptations so that these models can achieve better results.

5.2 Research Methodology

Our methodology is divided into four phases: A, B, C and D, as illustrated in Figure 20. Each phase is detailed in the next subsections.

5.2.1 Phase A - Dataset Fusion

In the first phase, we selected five datasets containing aggressive content memes: Facebook Harmful Meme [24]; MultiOFF [127]; Sexist Advertisement Database [29]; Harm-C and Harm-P [110]. These datasets were selected due to three reasons: 1) their use in previous research; 2) they have two clear class labels (0 for non-aggressive and 1 for aggressive memes); and 3) ease access, since they are publicly available. Details about the chosen datasets and their construction are presented in Section 5.3.

Once the five datasets have been selected, they were combined to compose only one dataset. Here, however, we used only the test splits of each individual dataset to compose the large one. This was done to try to reduce data overlap between these datasets and those used to train the Generative Models. If any meme dataset has been used to train this models, it is expected that its training split has been used only. In addition, since the next phase of our methodology involves manual annotation of memes from this large dataset into three classes, which is a time-consuming process, we randomly selected n memes instances from each test split, being $n/2$ memes with Label=1–aggressive, and $n/2$ with Label=0–non-aggressive.

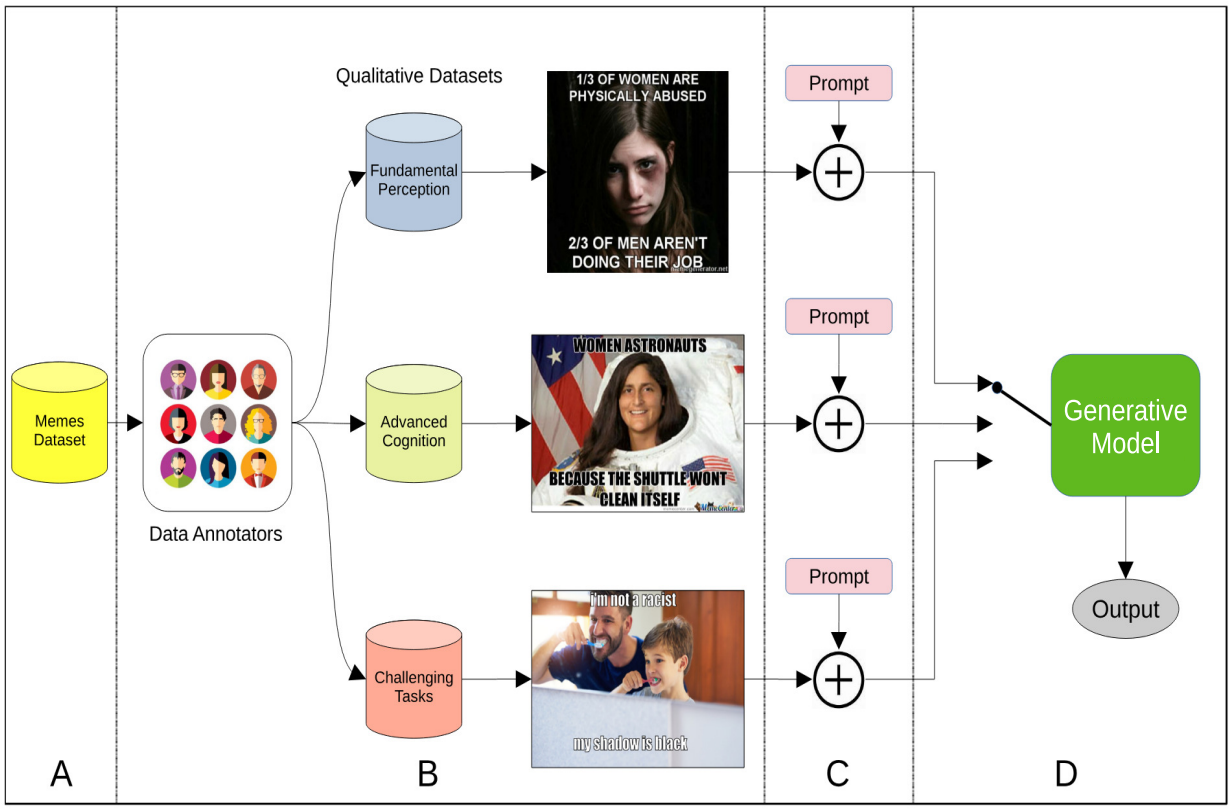

Figure 20 – The methodology conducted in this paper. First (A), several memes datasets are combined. Then (B), the grouped dataset is divided into three different datasets, each with different levels of multimodality reasoning to perceive aggressiveness, according to manual annotation. In sequence (C), a prompt is added to each meme instance. Finally (D), each meme and its corresponding prompt are presented to a Generative Model, which will generate an output Label=1 or Label=0, whether the meme contains aggressive content or not. Source: Author

Considering these instances defined, we are ready to start the next phase of our methodology, which is explained in the next subsection.

5.2.2 Phase B - Levels of Multimodality Reasoning to Perceive Aggressive Content in Memes

In this phase, the aggressive instances obtained in the previous phase were presented to a group of annotators who were responsible for assigning them to one of the three classes representing different levels of multimodality reasoning to perceive aggressive content. Each annotator decided which class to assign to each instance. At this point, we introduce the concept of Multimodality Reasoning for

the Presence Perception of Aggressive Content. Our hypothesis is that the presence of aggressive content in a meme is perceived differently by an individual, depending on the way in which this content is exposed in the meme. Based on this hypothesis, we propose the following three levels/classes of multimodality reasoning:

1. **Fundamental Perception (FP):** Memes in this category have both images and text that clearly show aggressive content, i.e. each modality separately shows aggressive tones, as illustrated in the FP example shown in Figure 20. Machine learning models should easily identify these memes as aggressive.

2. **Advanced Cognition (AC):** Memes here have one modality (either image or text) with clear aggressive content, while the other part does not, as can be observed in the AC instance shown in Figure 20. In this example, the text of the meme clearly indicates aggressive content. However, it is important to mention that the analysis of both modalities reinforces the aggression message. Therefore, ignoring one modality can cause errors in identifying aggression.

3. **Challenging Tasks (CT):** Memes in this category have neither the image nor text showing aggression when looked at separately. However, when combined, they show aggressive content, as the CT meme illustrated in Figure 20. This requires a full analysis of both parts together. Therefore, a machine learning model must rely heavily on multimodal reasoning so as to correctly identify these memes as aggressive, since memes often do not have a clear connection between visual and textual elements.

At the end of this phase, three different datasets were created from the large original dataset obtained in the first phase of our methodology. We call them Qualitative Datasets: FP, AC, and CT. Their annotation protocol is described in Section 5.3.3. It is worth noting that this second phase is essential to answer our Research Question 1. Our proposal is that by creating this subdivision we will be able to better evaluate the performance of the models in the task.

5.2.3 Phase C - Prompt Integration

A prompt is a mix of natural language instructions and media content, such as an image, designed to give models all the information they need to understand one request correctly. Prompts allow people without deep knowledge in machine learning to interact with and benefit from Generative Models. For complex tasks, such as detecting aggressive content in memes, creating effective prompts is not simple. It is necessary knowledge, experience, and a lot of experimentation to understand how a model behaves and guide it to achieve the desired results [134].

Figure 20 shows that we propose to integrate prompts with each meme in the phase C of our methodology. The prompt is presented in text format and, together with the meme itself, forms the input for the model. One important point is that we do not need to extract the text present in the meme image, since Generative Models can obtain this information directly from the meme. In this paper, we created two different prompts, both using the Zero-Shot approach [135, 52, 134]. The first has few words and directly describes our request, while the second presents a more complex content, incorporating a series of features that inform the model different aspects that should be taken into account during the classification process. We list the features incorporated into the second prompt model below.

- **Comprehensive Analysis:** The prompt guides the model to consider various aspects of the meme, including textual content, tone, visual elements, cultural context, and target audience.
- **Identification of Aggressive Content:** It provides specific criteria for identifying aggressive content, such as offensive language, stereotypes, discrimination, and attacks on individuals or groups.
- **Purpose Evaluation:** It prompts the model to consider the purpose of the meme, whether it is meant for entertainment or if it is intended to attack someone or something.
- **Clear Instructions for Categorization:** After analyzing the meme, the prompt instructs the model to categorize it as either aggressive (Label=1) or non-

aggressive (Label=0), providing clarity on how to label the content based on the analysis.

- **Emphasis on Importance:** The prompt highlights the significance of the task, emphasizing its contribution to understanding and filtering content for appropriateness.

These characteristics form the basis for answering our second research question, which is how to best optimize prompts to improve the ability of the models to effectively detect aggressive content within memes. In our experiments, the two prompts were used to evaluate the models' performance. In Section 5.3, we present the two prompts created to perform this task. Finally, after defining the prompts, the next phase involves using Generative Models, explained in the next subsection.

5.2.4 Phase D - Generative Models

Generative models are neural networks designed to create new data, such as text or images, based on patterns learned from existing data. Recent advances in these models have enabled them to perform complex tasks like generating content, reasoning, and understanding language. Below, we describe three key models: GPT-4V, Gemini, and LLaVA.

- The GPT-4V is an enhanced version of the GPT-4 model, developed by OpenAI. It can process both text and images, making it useful for tasks like image descriptions, interpreting graphs, and integrating visual and text-based information. GPT-4V maintains strong text generation capabilities while adding visual reasoning.

- Gemini, developed by Google DeepMind, is designed to excel at language tasks and also integrates data from different sources. It uses reinforcement learning and supervised learning to improve its performance in more complex tasks. Gemini is similar to models like GPT-4V but focuses more on integrating information from various inputs.

- The LLaVA (Large Language and Vision Assistant) model combines vision and language processing. It is built to handle tasks that require both image analysis and text interpretation. LLaVA aligns visual and textual information effectively, making it strong in tasks that involve visual context, like image descriptions.

Therefore, in the final phase of our methodology, all sets of images and prompts from each qualitative dataset are used to feed the three Generative Model. In our experiments, we calculated the following metrics for each model: Accuracy, Precision, Recall, and F1-Score. This allowed us to compare their performances in each qualitative dataset.

5.3 Experiments

This section details the experiments performed and shows the results. The main objective is to evaluate the behavior of the models on perceiving aggressive content in memes by varying the level of multimodality reasoning demanded. These levels are represented by the qualitative datasets. In Subsection 5.3.1, we detail the original meme datasets chosen to generate the qualitative datasets, while Subsection 5.3.2 provides information related to the composition and diversity of these datasets. In Subsection 5.3.3, we describe the process used by the annotators to assign each meme to a qualitative dataset. In Subsection 5.3.4, we present the prompts used in our experiments. Then, in Subsection 5.3.5, we show the results of the experiments.

5.3.1 Original Memes Datasets

During the experiments, five meme datasets were used to create the qualitative datasets. They were chosen because they are used in many works, such as those described throughout this thesis. These datasets are as follows:

- Facebook Harmful Meme (FBHM) [24]: This dataset has over 10,000 memes

showing different types of aggression against legally protected groups. Each meme is assigned to just one attack category, making it a binary classification task. See Figure 21 (a) for an example.

- MultiOFF [127]: It is composed of 743 memes whose content is devoted to the 2016 United States presidential election. These memes are classified as either offensive or non-offensive. An example is shown in Figure 21 (b).

- Sexist Advertisement Database (SAD) [29]: It is composed of 800 memes with sexist (and non-sexist) content, labeled based on visual and/or textual aspects. Figure 21 (c) shows one example of a meme from this dateset.

- Harm-C [110]: It contains 4793 memes related to COVID-19. These memes are classified as either offensive or non-offensive. Figure 21 (d) shows one example.

- Harm-P [110]: This dataset contains 5258 memes also focused on the United States politics. These memes are classified as either offensive or non-offensive. One example of an instance from this dataset is shown in Figure 21 (e).

(a) (b) (c) (d) (e)

Figure 21 – A sample of aggressive memes extracted from the (a) FBHM, (b) MultiOFF, (c) SAD, (d) Harm-C, and (e) Harm-P datasets respectively. Source: Author.

5.3.2 Datasets Composition and Diversity

It is important to observe that, although FBHM, MultiOFF, SAD, Harm-C, and Harm-P provide useful resources, their scope, as well as the scope of other memes datasets, is often narrow and focused on specific topics like politics, public health, or sexism. Consequently, this leaves significant gaps in other important areas. In this subsection, we present a more detailed analysis of the datasets used in this work in terms of sources, diversity and potential biases.

Figure 22 illustrates the proportion of aggressive and non-aggressive content across the five analyzed meme datasets. It is possible to observe that, as expected, the datasets are unbalanced. However, aggressive memes constitute the majority class, accounting for 59.9% of the total content. This higher proportion of aggressive memes is probably the result of bias during the organization of the dataset, given that, when creating a dataset of aggressive and non-aggressive memes, it is natural that more focus is given to collecting aggressive instances. Therefore, the higher proportion of aggressive memes may not reflect the reality observed on social media.

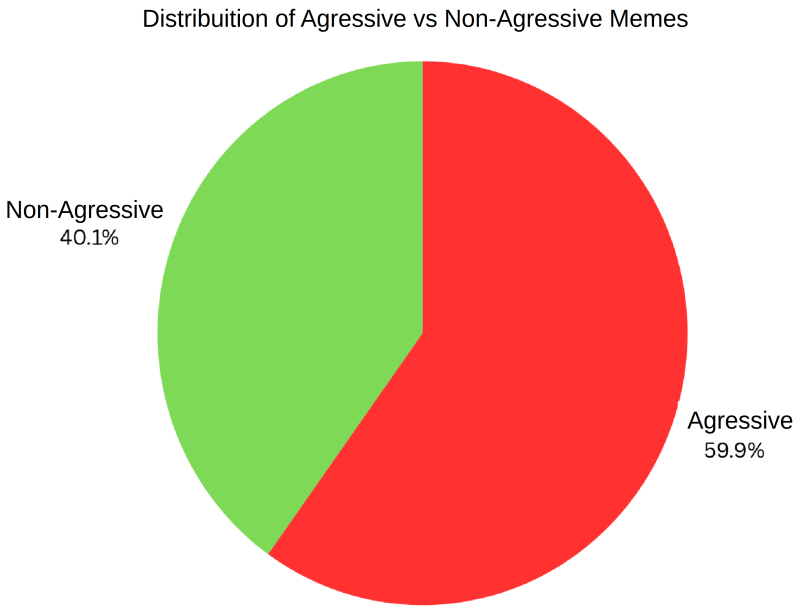

Figure 22 – Proportion of aggressive and non-aggressive memes across the analyzed meme datasets.

In addition, these aggressive memes include various types of aggressive or harmful themes, often targeting specific groups or individuals. Figure 23 shows the distribution of meme content between different categories. The largest portion is

devoted to Politics, representing 27.8% of the total of aggressive memes, reflecting a significant focus on political themes. Then, the second highest represented theme is COVID-19-related, representing 22.2% of the datasets, highlighting the impact of the pandemic on meme culture. Next, three themes related to harmful content stand out: Dehumanization, which comprises 18.5% and includes memes that portray individuals or groups as less than human; Aggression against groups, also at 18.5%, encompassing memes targeting specific groups with hostility; and Other Harmful Content, totaling 9.3% of memes whose harmful content does not fit into the previous two subcategories. Finally, Sexism is the least represented category, with 3.7%, composed by memes that perpetuate sexist ideas or stereotypes.

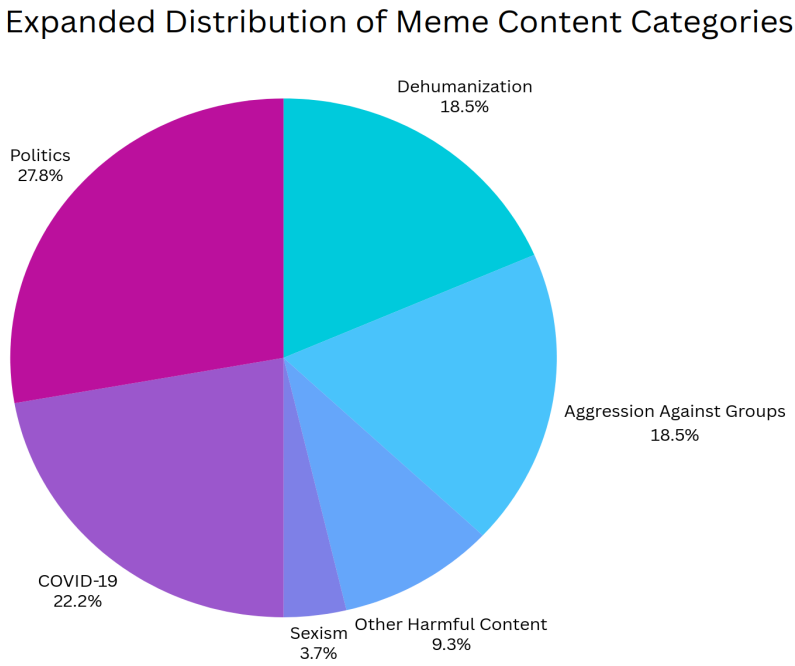

Figure 23 – Considering the expanded classification of aggressive meme content, this chart illustrates the proportional distribution across six distinct categories.

This distribution provides insights into the thematic focus of meme datasets and their potential societal implications. It also highlights the prevalence of potentially concerns due to low diversity content within memes datasets in general. Moreover, the quality of the existing datasets is often affected by issues such as cultural bias, under-representation of certain groups or contexts, and overly simplistic classification methods (e.g., labeling content as offensive or not). These limitations can result in biased analyses, reducing the effectiveness and applicability of models

trained on these datasets.

This also accentuates the urgent need to increase both the quantity and quality of datasets in this domain. It is fundamental to incorporate perspectives from different cultures, languages, and social contexts. In addition, improving annotation processes to address inherent biases, and transparently documenting dataset limitations are critical steps. These efforts are necessary to develop more reliable, ethical, and effective systems for analyzing and classifying aggressive memes.

5.3.3 Qualitative Datasets Process of Annotation

From the test split of each one of the five datasets described in the previous subsections, we randomly selected 60 memes labeled 1 and 60 memes labeled 0. Therefore, the fused dataset obtained in the phase A of our methodology was composed of 600 instances. Then, a group of 8 annotators reviewed each aggressive meme (300 instances) and were asked to assign to them one of the three classes of multimodality reasoning to perceive aggressive content in memes, defined in our protocol, based on the criteria defined in subsection 5.2.2. The class assigned for each meme was determined by majority voting among the annotators.

Table 10 – Class-wise data distribution of each original dataset obtained as a result of the manual annotation process. Columns show how each original meme dataset contributed to the qualitative datasets in the rows. For example, the 60 memes randomly chosen from the FBMH dataset resulted in 17 memes for FP, 32 for AC, and 11 for CT.

	FBHM	**MultiOFF**	**SAD**	**Harm-C**	**Harm-P**	Total
FP	17	10	19	19	9	74
AC	32	44	38	27	39	180
CT	11	6	3	14	12	46
Total	60	60	60	60	60	**300**

Table 10 depicts the class-wise data distribution of each original dataset obtained after the manual annotation process. In addition, Figure 24 summarizes the proportion of aggressive instances assigned to each qualitative dataset. Some observations can be made from this table and figure.

- Most memes (60%) belong to the AC category, meaning aggressive content

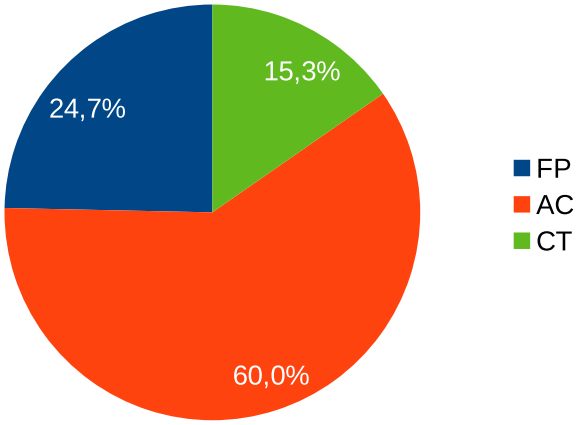

Figure 24 – Data distribution of the 300 aggressive memes obtained after the classification conducted by the annotators into three qualitative datasets (FP, AC, and CT). Source: Author

may be observed in only one modality (image or text). Although not the most challenging group, this class represents a difficult classification task, since the models need to consider both modalities but also need to be able to focus their attention to the modality providing the aggressive message.

- The FP category represents 24.7% of the memes, where aggressive content is clearly visible in both modalities. As previously mentioned, these instances are expected to be easier for models to identify.

- Only 15.3% of the memes fall into the CT category, the most challenging group, since neither the image nor the text of the meme in this group show aggression when considered separately, but they do together. This category requires from machine learning models a comprehensive analysis of both modalities to detect aggression content.

Analyzing these results, it can be concluded that memes demanding tackling both text and image are the majority, precisely 75% combining AC and CT. This allow us to confirm that the analysis of the content of a meme must take into account both modalities to allow capturing the original meaning of the meme's

content. On the other hand, the vast majority of memes fall into the AC group, whose message is clearly conveyed and generate immediate understanding when observing one of the two modalities individually. It is also important to reinforce that the CT memes are the minority group. Therefore, based on the results provided by our annotation process, only in less than 16% of memes the models really need to be able to tackle multimodal reasoning so as to classify the meme as aggressive. The next subsections present the characteristics defined and incorporated into the prompts during our experiments, and the results attained.

5.3.4 Prompts

In Subsection 5.2.3, we explained that two prompts were developed following the Zero-Shot method. The difference between them is the amount of information transmitted to the model. Zero-Shot was chosen due to its efficiency in addressing several tasks, benefits in adaptability, generalization and reduced dependence on labeled datasets. The two prompts used in our experiments are shown below.

- **First Prompt:** *Please, analyze the meme image provided. After your analysis, categorize the meme according to its content. Return ONLY THE TEXT 'Label=1' if the meme is aggressive and ONLY THE TEXT 'Label=0' if it is not aggressive.*
- **Second Prompt:** *Please, analyze the meme image provided, consider the text content, look at the words or phrases in the overlaid text. Check for offensive language, insults, slander, or any language that could be considered aggressive. Consider the tone of the text. Is it sarcastic, humorous, or does it provoke a negative emotional response? The tone can help determine whether the meme is harmless or aggressive. Take into account the cultural context of the meme. What may be considered harmless in one culture could be offensive in another. Consider the meme's target audience and the culture in which it exists. Evaluate the visual image. Is the image itself offensive or aggressive? Does it depict situations or symbols that might be interpreted as aggressive or disrespectful? Check if the meme contains stereotypes or discrimination based on race, gender, religion, sexual orientation, or any other characteristic. This can make the*

meme aggressive. Consider whether the meme attacks or ridicules a specific individual, group, or idea in a way that could cause harm or offense. Ask yourself what the purpose of the meme is. Is it meant to entertain, or is it attacking someone or something? The purpose can help determine if the meme is aggressive. After your analysis, categorize the meme according to its content. Return ONLY THE TEXT 'Label=1' if the meme is aggressive and ONLY THE TEXT 'Label=0' if it is not aggressive. Your task is important as it contributes to understanding and filtering content for appropriateness.

The second prompt for analyzing memes in an MLLM was developed through a structured process involving reverse engineering, iterative testing, and refinement to ensure comprehensive meme classification. The approach began with data collection and reverse engineering, where a diverse dataset of memes, including aggressive and non-aggressive examples, was analyzed. This dataset covered a variety of meme structures, tonal variations, cultural contexts, and content categories, allowing for a detailed breakdown of common linguistic and visual patterns associated with harmful content. By deconstructing these memes, key features contributing to aggression were identified, forming the basis for the criteria outlined in the prompt. Following this analysis, the core analytical components were established, ensuring that the MLLM evaluated memes holistically rather than relying solely on textual cues. These components included textual content analysis to extract and assess language, tone analysis to differentiate between sarcasm and genuine aggression, cultural sensitivity to account for regional variations in meme interpretation, visual content inspection to identify offensive imagery, stereotype and discrimination detection to flag implicit biases, and intent assessment to distinguish between humor and targeted harassment. Structuring the prompt around these six key areas provided the MLLM with a systematic and robust framework for meme analysis. Once the first draft of the prompt was created, an iterative refinement process was initiated. The initial prompt was tested on an MLLM. Several issues emerged during testing, leading to multiple refinements. Early iterations revealed ambiguity in responses, as the model sometimes provided explanations instead of binary classifications. This issue was resolved by explicitly

instructing the model to return only Label=1 for aggressive memes and Label=0 for non-aggressive memes. Additionally, the model struggled with sarcasm and implicit aggression, prompting the inclusion of a directive to evaluate tone, humor, and emotional responses. Similarly, cultural context misinterpretations were addressed by specifying that the model should consider the meme's target audience and cultural relevance. Another refinement involved improving the model's ability to detect image-based aggression, as some harmful memes contained offensive symbols without explicit textual cues. To address this, the prompt was updated to ensure the evaluation of visual elements alongside textual analysis. Finally, the model occasionally misclassified satirical memes as aggressive, leading to the incorporation of an intent-based evaluation criterion, ensuring the meme's purpose—whether meant for entertainment or harm—was considered in classification. Through this iterative testing and refinement process, the final prompt was optimized to enhance clarity, precision, and adaptability while maintaining a structured binary classification output. The development of this prompt through reverse engineering ensured that the MLLM could effectively analyze memes by considering text, visuals, cultural awareness, and intent, providing a reliable and scalable approach for automated content moderation.

5.3.5 Experiments Results

The experiments were conducted according to the methodology described in Figure 20. The results attained are presented in this section grouped by each one of the three models investigated. It is important to mention that a discussion related to these results is presented in the next section.

In Table 11 we see a summary of the performances reached by the GPT-4V model. As expected, the best results were obtained using the second prompt (2^{nd} P.) in all three qualitative datasets. The superiority of the 2^{nd} P. was especially evident in FP, where the F1-Score increased from 0.58 to 0.75. Even in the most difficult dataset (CT), the performance improved, although not as much, with the use of the 2^{nd} P. This indicates that using more detailed prompts helps the model to

better deal with the complexity and nuances of perceiving aggressiveness in memes. Considering the qualitative datasets, the results indicate an order relation between the performances attained: the best results were obtained in FP (F1-Score 0.75), the second best in AC (F1-Score 0.65) and the worst values were attained in CT (F1-Score 0.57). This behavior confirms our hypothesis that the difficulty on detecting aggressive memes increases as the level of multimodality reasoning necessary to perceive aggressive content also increases.

Table 11 – Results attained by the GPT-4V model using the 1^{th} and 2^{nd} prompts on the Qualitative Datasets.

Model	FP		AC		CT	
GPT-4V	1^{th} P.	2^{nd} P.	1^{th} P.	2^{nd} P.	1^{th} P.	2^{nd} P.
Accuracy	0.58	0.74	0.60	0.64	0.62	0.58
Precision	0.69	0.73	0.64	0.63	0.67	0.59
Recall	0.50	0.77	0.47	0.66	0.45	0.56
F1-Score	0.58	0.75	0.54	0.65	0.54	0.57

Table 12 – Results attained by the LLaVA model using the 1^{th} and 2^{nd} prompts on the Qualitative Datasets.

Model	FP		AC		CT	
LLaVA	1^{th} P.	2^{nd} P.	1^{th} P.	2^{nd} P.	1^{th} P.	2^{nd} P.
Accuracy	0.62	0.60	0.50	0.54	0.45	0.52
Precision	0.62	0.58	0.50	0.53	0.44	0.51
Recall	0.63	0.70	0.48	0.64	0.32	0.63
F1-Score	0.62	0.63	0.50	0.58	0.37	0.56

The superiority of the 2^{nd} P. is also observed when analyzing the LLaVA, shown in Table 12. Here, however, the difference was smaller, with the CT dataset presenting the highest difference: F1-Score increased from 0.37 to 0.56. Again, the same order of classification difficulty among the three qualitative datasets was observed. When comparing GPT-4V to LLaVA, it is possible to see that GPT-4V outperformed LLaVA, especially when using the 2^{nd} P, across all datasets, with GPT-4V attaining higher accuracy and F1-Score. On the other hand, the difference between the performances presented by both models decreases as the level of multimodality increases. The smallest difference is in CT, where both models attained performances slightly better than the performance of a method that randomly guesses the instance label (F1-Score of 0.57 and 0.56 for GPT-4V and LLaVA respectively).

This suggests that these models do not handle successfully the complexity and subtlety of aggressive memes when more multimodality reasoning is necessary.

Table 13 – Results attained by the Gemini model using the 1^{th} and 2^{nd} prompts on the Qualitative Datasets.

Model	FP		AC		CT	
Gemini	1^{th} P.	2^{nd} P.	1^{th} P.	2^{nd} P.	1^{th} P.	2^{nd} P.
Accuracy	0.73	0.70	0.60	0.60	0.57	0.57
Precision	0.69	0.64	0.58	0.57	0.56	0.55
Recall	0.85	0.89	0.62	0.74	0.63	0.76
F1-Score	0.76	0.75	0.60	0.65	0.59	0.64

Finally, the results reached by Gemini followed a pattern slightly different, as shown in Table 13. First, the 2^{nd} P. did not help Gemini to improve its results so effectively as it was observed in the two previous models. The accuracy rates attained by the models using the prompts were similar, while the F1-Score improved only 0.05 in AC and CT. However, Gemini performed best overall, outperforming both LLaVA and GPT-4V in most cases, especially when comparing the F1-Score rates. This suggests that Gemini may be most effective in detecting different aggressive memes and that it is necessary to design a more detailed prompt to help it to improve its detection capacity. Finally, even though the F1-Score in CT was better, the lowest classification rates were again obtained in this dataset. In the next section, the obtained results are discussed in details, as well as ethical issues.

5.4 Discussion

The experiments conducted in this chapter revealed several important observations. When we consider the overall performance, Gemini consistently showed the best results among the three models, across all datasets and with both types of prompts. Gemini achieved the highest F1-score, especially on the FP dataset (0.76 using the first prompt and 0.75 with the second prompt). GPT-4V performed very closely to Gemini, particularly on the FP dataset, where it also achieved F1-Score of 0.75 with the second prompt. However, its performance on the other datasets, especially AC, was slightly lower. LLaVA attained the lowest overall performance among the three models, although it showed improvements when using the second

prompt, as detailed in Figure 25.

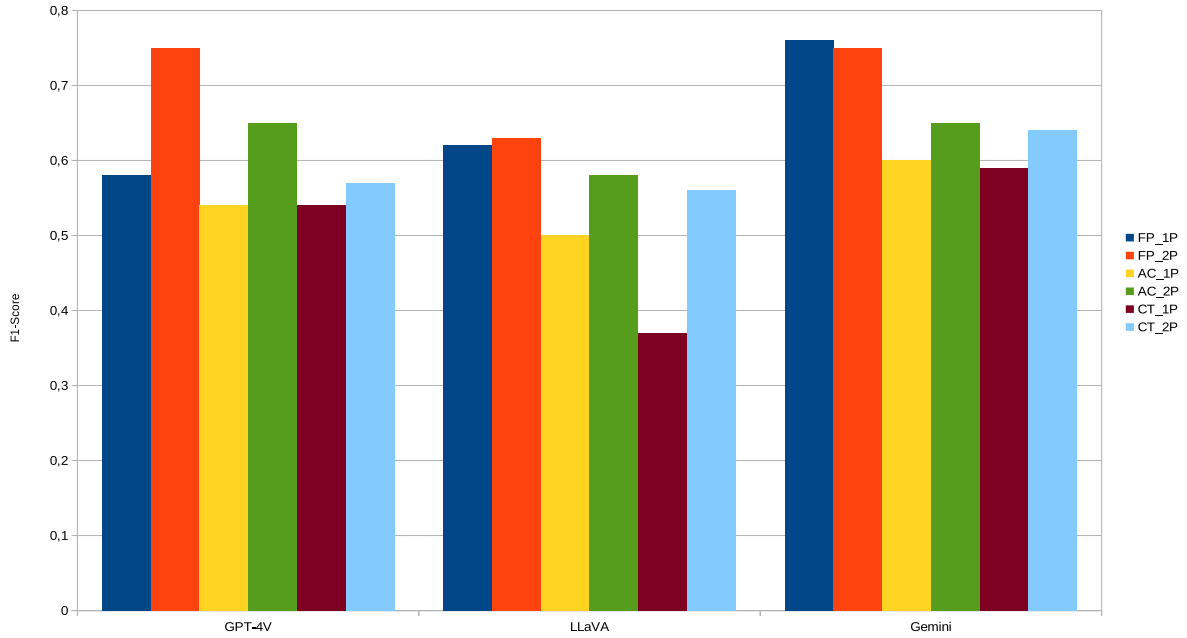

Figure 25 – Bar chart comparing the F1-Scores of GPT-4V, LLaVA, and Gemini across the three datasets (FP, AC, CT) with two different prompts (1^{th} P. and 2^{nd} P.). The chart illustrates how each model's performance varies depending on the dataset and the prompt used. Source: Author

Considering the impact of the prompts, the second prompt helped to significantly improve the performance of all models compared to the first prompt, indicating that providing more detailed instructions helps models to better identify aggressive content. Despite being less impacted than the other two models, Gemini benefited from the second prompt, with improvements in recall and F1-Score, particularly on CT. GPT-4V presented improvements with the second prompt especially on FP, where its F1-Score increased from 0.58 to 0.75. Finally, LLaVA obtained the most substantial improvements with the second prompt on AC, with an increase in F1-Score from 0.50 to 0.58. These results are detailed in Figure 26.

Finally, when we consider the performance by varying the dataset, all results confirm our hypothesis that the difficulty on detecting aggressive memes increases according to the level of multimodality reasoning necessary to perceive aggressive content. It is detailed in Figure 27 that on FP, where both the image and text of the memes are aggressive, Gemini and GPT-4V showed F1-Scores superior to 0.70. This is a very interesting result, especially taking into account that no training

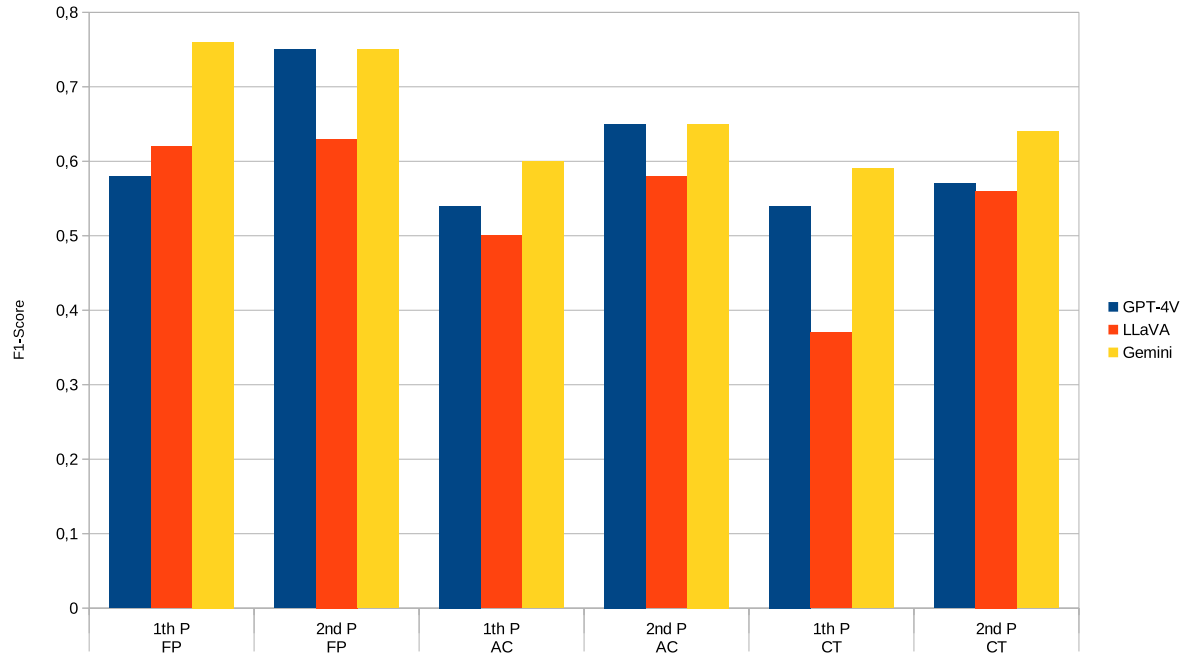

Figure 26 – Bar chart illustrating the F1-Score of GPT-4V, LLaVA, and Gemini on FP, AC, CT datasets using the 1^{th} P. and 2^{nd} P. The chart compares the performance of each model on the same datasets under different prompts, highlighting the effectiveness of the second prompt (2^{nd} P.) in improving the models' ability to detect aggressive content. The Gemini model consistently outperforms the other models across all datasets and prompts. Source: Author

was conducted. In its turn, on AC, where only one element (image or text) is aggressive, all models reduced their performance, but Gemini and GPT-4V were able to reach F1-Scores superior to 0.60 using the second prompt. However, on the most challenging dataset (CT), where the combination of image and text results in aggressiveness, except for Gemini which attained F1-Score of 0.64 using the second prompt, the models obtained F-Scores slightly superior to 0.50, which is similar to random guess.

Therefore, based on our results, all three MLLMs demonstrated the ability to detect aggressive memes using the Zero-Shot approach, despite showing difficulty in handling complex multimodal content. In order to better understand the extent of their ability, we compare their results to the results obtained by a multimodal model trained using transfer learning. This comparison is presented in the next subsection.

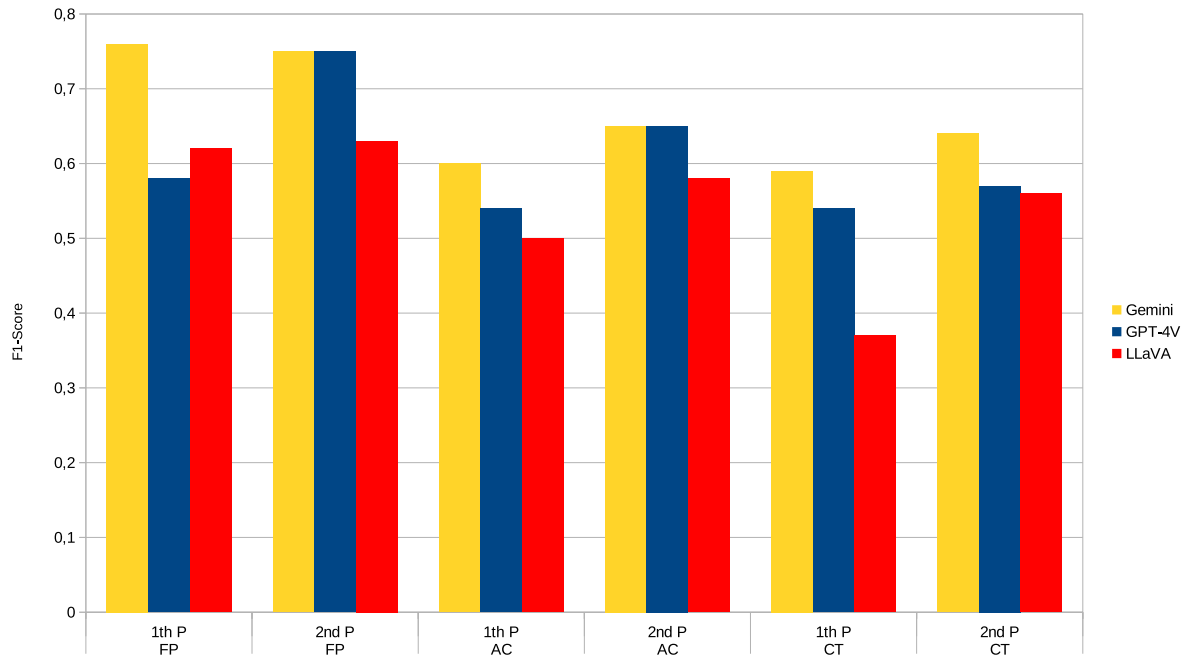

Figure 27 – The bar chart shows the F1-Scores of Gemini, GPT-4V, and LLaVA across the three datasets FP, AC, CT using the two different prompts (1^{th} P. and 2^{nd} P.). Each bar represents the performance of a model on a specific dataset and prompt, highlighting the variations in model effectiveness. Source: Author

5.4.1 Comparing MLLMs with a Specialized Multimodal Model

We compare the results achieved by the MLLMs to those reached by the VisualBERT-CPB, described in the previous chapter. It is important to mention that this model was trained in the training split of each one of the five original datasets used to compose our Qualitative Datasets: FP, AC and CT. Therefore, five trained models were obtained.

To try to provide a fair comparison with the Generative Models, the following procedure was performed. Each VisualBERT-CPB was used to make predictions on the test splits corresponding to data from the Qualitative Datasets. For instance, the model trained on the FBHM dataset was evaluated in three distinct test splits. The first associated with FP, consisting of 34 memes (17 labeled as 1 and 17 labeled as 0); the second split is linked to AC, containing 64 memes; and the third split is linked to CT, with 22 memes. This approach was applied consistently across the other four models linked to different datasets. Then, the F1-Score was calculated considering the results of the five models on each Qualitative Dataset. Moreover, to simplify the comparison, only the results obtained by Gemini, utilizing the 2^{nd}

prompt, which demonstrated the best performance among the Generative Models, are reported here. The comparison is summarized in Table 14.

Table 14 – Comparing F1-Scores attained by Gemini (using the second prompt) and VisualBERT-CPB across the three Qualitative Datasets: FP, AC, and CT.

F1-Score	FP	AC	CT
Gemini - 2^{nd} Prompt	0.75	0.65	0.64
VisualBERT-CPB	0.79	0.64	0.82

These results reinforce the fact that Generative Models are viable for the task of meme classification via Zero-Shot. Gemini using the second prompt exhibited competitive performance, particularly on the AC dataset. In terms of FP, VisualBERT-CPB was only slightly superior. On the other hand, VisualBERT-CPB significantly outperformed Gemini on the CT dataset. These findings suggest that MLLMs can be effective in meme classification but also highlight the need to explore more advanced prompt techniques to allow the multimodal information between the two modals of the memes to be fully explored and used. Refining prompts or developing new prompt engineering approaches could potentially increase the accuracy and robustness of MLLMs in future tasks, leading the methods to be capable of truly understanding multimodality.

5.5 Limitations and ethical issues

The use of Multimodal Large Language Models (MLLMs) for content moderation presents significant challenges, particularly in detecting aggressive memes. While these models offer scalability and automation, their application raises concerns regarding ethical implications, fairness, and technical reliability. One of the primary concerns in deploying MLLMs for content moderation is the risk of bias and censorship. The presence of false positives, where non-aggressive content is flagged as harmful, and false negatives, where harmful content goes undetected, can lead to disproportionate restrictions on free speech or the unchecked spread of offensive material. Automated systems may mistakenly flag satirical, artistic, or politically critical memes as aggressive, leading to unjustified content removal. Additionally, memes often rely on coded language or cultural nuances that are difficult for mod-

els to interpret, allowing harmful content to bypass detection. MLLMs also inherit biases present in their training datasets, leading to discriminatory moderation that disproportionately impacts certain groups or ideologies, perpetuating systemic inequalities, particularly in politically or culturally sensitive contexts.

The reliance on opaque moderation algorithms raises questions about transparency and accountability. Users often have limited insight into why content is removed, and the lack of a clear appeals process can result in arbitrary or unfair decisions. Given the increasing use of MLLMs in moderation, there is a growing need for ethical guidelines that prioritize fairness, explainability, and user rights. Despite their advanced multimodal capabilities, MLLMs still face significant limitations in reasoning about complex or highly contextualized memes. Experimental results reinforce these challenges, with the best-performing model, Gemini, achieving only a 0.64 F1-score on the CT dataset. This low score suggests that MLLMs struggle with nuanced multimodal reasoning, particularly in cases where the text and image components must be interpreted together to determine aggression, the meme relies on sarcasm, cultural references, or implicit messages that cannot be easily mapped to explicit rules, or the model encounters ambiguous or edge-case content that even human moderators may struggle to classify.

A key difficulty in evaluating MLLMs is the lack of a standardized benchmark that accurately captures multimodal reasoning complexity. Many existing datasets focus on explicitly aggressive language, but memes often require higher-order inference skills, including an understanding of historical references, humor, and social dynamics. Without structured evaluation metrics, it is difficult to determine whether a model's failure is due to a lack of knowledge, reasoning errors, or dataset biases. Given these limitations, it is evident that MLLMs alone cannot ensure fully reliable moderation. Over-reliance on automated systems introduces operational risks, including the misclassification of sensitive content, as false positives and false negatives can result in legal and reputational risks for platforms implementing MLLM-based moderation. Additionally, users may intentionally design memes that exploit model weaknesses, bypassing moderation rules through subtle modifications in text or imagery. There is also a trade-off between scalability and accuracy, as while MLLMs can process large volumes of content efficiently, their inability to

fully understand context necessitates a hybrid approach that integrates human moderation.

To mitigate these risks, human-AI collaboration in content moderation is essential. A hybrid model, where MLLMs assist human moderators rather than replace them, can help ensure more accurate and context-aware decisions. Transparent reporting mechanisms, appeals processes, and continuous model evaluation are crucial to maintaining trust and fairness in automated content moderation. While MLLMs present great potential for detecting aggressive content in complex formats like memes, their deployment must be carefully managed to avoid unintended censorship, bias, and misclassification. Achieving a balance between safety and freedom of expression requires not only technological advancements but also ethical guidelines and collaboration among platforms, governments, and civil society.

5.6 Final Remarks

This study investigated the ability of three leading Generative Models, specifically, GPT-4V, Gemini, and LAaVA, to identify aggressive content in memes. To achieve this, we selected a diverse dataset of memes collected from five different sources. These memes were then categorized into three Qualitative Datasets based on their levels of multimodality reasoning necessary to perceive aggressive content. The Qualitative Datasets were obtained by manual annotation conducted by 8 different annotators. We developed two prompts using the Zero-Shot technique—the first being simpler and the second being more complex, incorporating key features that aided the models in the classification process. Our results demonstrate the significant potential of Generative Models in detecting aggressive content in memes. However, all three methods showed difficulty in dealing with high levels of multimodality reasoning, since they reached low classification rates on the most challenging Qualitative Dataset.

Moreover, we acknowledge the inherent limitations of relying solely on automated detection of aggressive content in memes, and advocate a hybrid approach. This approach could combine the efficiency of Generative Models with the judg-

ment of human experts, ensuring higher transparency, accountability, and trust in the detection process. This research not only highlights the feasibility of using Generative Models for this challenging task, but also emphasizes the importance of a careful implementation to address ethical issues.

The scope of this study was intentionally limited to prioritize the establishment of a robust methodology and the identification of key prompt characteristics that allow the models to be used at different levels of meme multimodality. Similarly, while we evaluated three top-performing models in their respective categories, our focus was not on creating a definitive classification of all existing models. Rather, we focused on developing a methodology that can be used in future classification efforts and adaptable to multiple tasks beyond aggressive meme detection.

The next chapter presents the final conclusions of this thesis.

6 Conclusion

This thesis focused on expanding knowledge and improving the ability to detect harmful content in memes, a complex but critical task in the online environment today. Our initial study, "Detecting Hate Speech in Memes: a Review" provided a comprehensive overview of the latest methods, introducing a specific taxonomy for hateful meme detection techniques. This structure enables consistent evaluation of methods and results, identifying strengths and gaps in existing approaches. Additionally, the critical analysis offers a solid foundation for future research by highlighting emerging trends and challenges in hate speech detection within memes.

The second contribution, "Adding Compact Parameter Blocks to Multimodal Transformers to Detect Harmful Memes" sought to improve the accuracy of multimodal models by incorporating Compact Parameter Blocks (CPBs) into the Transformer architecture. This modification not only reduced the processing burden on attention mechanisms but also strengthened the model's capacity to identify nuanced multimodal signals within harmful memes, proving to be an effective strategy for enhancing detection accuracy.

Finally, the study "Exploring the Performance of Generative Models in Detecting Aggressive Content in Memes" explored the potential of Multimodal Large Language Models in detecting aggressive content in memes when the levels of multimodality reasoning necessary to classify the meme are varied. By employing prompt engineering, these models were optimized for multimodal analysis, demonstrating a competitive performance in identifying aggression in memes, especially those with moderate or low multimodal complexity. The research indicated that, compared to specialized models, Generative Models can be competitive in the task of detecting memes with aggressive content as long as the multimodality reasoning level required is not very high. Moreover, these results also show that it is possible to use these models in more refined and inclusive moderation tools.

Together, these contributions advance the state-of-the-art in harmful meme detection, providing a robust theoretical foundation, and introducing innovative practical approaches. These advancements are important to contribute in providing a

safer online environments while keeping the principle of freedom of expression.

6.1 Future Works

Future research can improve the detection of harmful memes by adding context and cultural understanding, making models more accurate for different audiences. Developing tools that distinguish between humor and harmful intent will help in detecting subtle aggression. Furthermore, creating real-time moderation tools using multimodal analysis could allow quicker responses to harmful content. Addressing issues like model transparency and data privacy will be essential as detection methods advance, with the aim of creating a safer and more inclusive online environment.

Our future research will be divided into two distinct areas of investigation. The first will concentrate on refining the efficacy of CPBs, encompassing a thorough examination of potential modifications to the fundamental CPB architecture, an assessment of the optimal place for the block within the original models, and an exploration of the potential benefits of incorporating multiple blocks, either with uniform or diverse attributes, into a single model.

The second area is focused on Generative Models and their performance in the context of harmful content detection within memes. Our research will investigate how to generalize the identified prompt characteristics to other prompting techniques for broader applicability across multiple models. Furthermore, we will refine the methodology to create Qualitative Datasets by expanding segmentation criteria, exploring different meme categories by considering groupings other than aggressiveness, and analyzing the model's behavior in more specific contexts.

References

[1] Statista. *Internet and social media users in the world 2024*. https://www.statista.com/statistics/617136/digital-population-worldwide/. Accessed: 2024-10-23. 2024.

[2] Arnav Arora et al. "Detecting harmful content on online platforms: what platforms need vs. where research efforts go". In: *ACM Computing Surveys* 56.3 (2023), pp. 1–17.

[3] Sian Brooke. ""Condescending, rude, assholes": Framing gender and hostility on stack overflow". In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 172–180. DOI: 10.18653/v1/W19-3519. URL: https://doi.org/10.18653/v1/W19-3519.

[4] Srecko Joksimovic et al. "Automated identification of verbally abusive behaviors in online discussions". In: (2019). DOI: 10.18653/v1/W19-3505. URL: https://doi.org/10.18653/v1/W19-3505.

[5] Hongzhan Lin et al. "Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models". In: *arXiv preprint arXiv:2401.132* (2024).

[6] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (Feb. 2019), pp. 423–443. DOI: 10.1109/tpami.2018.2798607. URL: https://doi.org/10.1109%2Ftpami.2018.2798607.

[7] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008. DOI: 10.48550/arXiv.1706.03762. URL: https://doi.org/10.48550/arXiv.1706.03762.

[8] Phillip Lippe et al. "A Multimodal Framework for the Detection of Hateful Memes". In: *arXiv preprint arXiv:2012.12871* (2020).

[9] Zhenzhong Lan et al. “Albert: A lite bert for self-supervised learning of language representations”. In: *arXiv preprint arXiv:1909.11942* (2019). DOI: 10.48550/arXiv.1909.11942. URL: https://doi.org/10.48550/arXiv.1909.11942.

[10] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. “Detecting Hate Speech in Multi-modal Memes”. In: *arXiv preprint arXiv:2012.14891* (2020).

[11] Niklas Muennighoff. “Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes”. In: *arXiv preprint arXiv:2012.07788* (2020).

[12] Vlad Sandulescu. “Detecting Hateful Memes Using a Multimodal Deep Ensemble”. In: *arXiv preprint arXiv:2012.13235* (2020).

[13] Gokul Karthik Kumar and Karthik Nanadakumar. “Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features”. In: *arXiv preprint arXiv:2210.05916* (2022). DOI: 10.48550/arXiv.2210.05916. URL: https://doi.org/10.48550/arXiv.2210.05916.

[14] OpenAI. *GPTV System Card*. OpenAI Website. Acesso em: 20/09/2024. 2024. URL: https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[15] Llava-Vl. *LLAVA-VL Official Website*. 2024. URL: https://llava-vl.github.io/.

[16] DeepMind. *Gemini: A Multi-modal Neural Network for Goal-directed Agents*. Tech. rep. DeepMind, 2022. URL: https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

[17] Google. *Google Gemini: New Features July 2024*. Accessed: 2024-07-28. 2024. URL: https://blog.google/products/gemini/google-gemini-new-features-july-2024/.

[18] Haotian Liu et al. “Visual Instruction Tuning”. In: *NeurIPS*. 2023.

[19] Paulo Cezar de Q Hermida and Eulanda M dos Santos. “Detecting hate speech in memes: a review”. In: *Artificial Intelligence Review* 56.11 (2023), pp. 12833–12851.

[20] Paulo Hermida and Eulanda M Dos Santos. “Adding compact parameter blocks to multimodal transformers to detect harmful memes”. In: *Engineering Applications of Artificial Intelligence* 137 (2024), p. 109136.

[21] Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.

[22] Dylan Norris. “Memes and Hate Speech: Challenges in Content Moderation”. In: *Journal of Digital Ethics* 12.3 (2020), pp. 245–261.

[23] Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press, 2018.

[24] Douwe Kiela et al. “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes”. In: *arXiv preprint arXiv:2005.04790* (2020).

[25] Homa Hosseinmardi et al. “Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network”. In: *International Conference on Social Informatics*. Springer. 2014, pp. 49–66.

[26] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. “Ex Machina: Personal Attacks Seen at Scale”. In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2017, pp. 1391–1399.

[27] Shardul Suryawanshi et al. “Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text”. In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. 2020, pp. 32–41.

[28] Paula Fortuna and Sérgio Nunes. “A Survey on Automatic Detection of Hate Speech in Text”. In: *ACM Computing Surveys (CSUR)* 51.4 (2018), pp. 1–30.

[29] Francesca Gasparini et al. “Multimodal Classification of Sexist Advertisements.” In: *ICETE (1)*. 2018, pp. 565–572.

[30] Pinkesh Badjatiya et al. “Deep Learning for Hate Speech Detection in Tweets”. In: *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee. 2017, pp. 759–760.

[31] Geoffrey E Hinton and Ruslan R Salakhutdinov. “Reducing the Dimensionality of Data with Neural Networks”. In: *Science*. Vol. 313. 5786. American Association for the Advancement of Science, 2006, pp. 504–507.

[32] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[33] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[34] Pascal Vincent et al. “Extracting and Composing Robust Features with Denoising Autoencoders”. In: *Proceedings of the 25th International Conference on Machine Learning (ICML)*. ACM. 2008, pp. 1096–1103.

[35] Andrew Ng. “Sparse Autoencoder”. In: *CS294A Lecture Notes* 72.2011 (2011), pp. 1–19.

[36] Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Representation Learning: A Review and New Perspectives”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Vol. 35. 8. IEEE, 2013, pp. 1798–1828.

[37] W. X. Zhao et al. “A survey of large language models”. In: *arXiv* 2303.18223 (2023).

[38] OpenAI. *GPT-4 technical report*. Tech. rep. 2023.

[39] OpenAI. *ChatGPT: A language model for conversational AI*. Tech. rep. Available: `https://www.openai.com/research/chatgpt`. OpenAI, 2023.

[40] W.-L. Chiang et al. “Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality”. In: (2023). Available: `https://vicuna.lmsys.org`.

[41] H. Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv* 2302.13971 (2023).

[42] A. Kirillov et al. “Segment anything”. In: *arXiv* 2304.02643 (2023).

[43] Y. Shen et al. “Aligning and prompting everything all at once for universal visual perception”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[44] H. Zhang et al. “Dino: Detr with improved denoising anchor boxes for end-to-end object detection”. In: *arXiv* 2203.03605 (2022).

[45] M. Oquab et al. “Dinov2: Learning robust visual features without supervision”. In: *arXiv* 2304.07193 (2023).

[46] Shukang Yin et al. “A survey on multimodal large language models”. In: *arXiv preprint arXiv:2306.13549* (2023).

[47] Mr D Murahari Reddy et al. “Dall-e: Creating images from text”. In: *UGC Care Group I Journal* 8.14 (2021), pp. 71–75.

[48] Tom B. Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. DOI: 10.48550/arXiv.2005.14165. URL: https://arxiv.org/abs/2005.14165.

[49] Laria Reynolds and Kyle McDonell. *Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm*. 2021. DOI: 10.48550/arXiv.2102.07350. arXiv: 2102.07350 [cs.CL]. URL: https://arxiv.org/abs/2102.07350.

[50] Rishi Bommasani et al. “On the Opportunities and Risks of Foundation Models”. In: *arXiv preprint arXiv:2108.07258*. 2021.

[51] Alec Radford et al. “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 2021, pp. 8748–8763.

[52] Pranab Sahoo et al. “A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications”. In: *arXiv preprint arXiv:2402.07927* (2024).

[53] Teresa Quintel and Carsten Ullrich. “Self-regulation of fundamental rights? The EU Code of Conduct on Hate Speech, related initiatives and beyond”. In: *Fundamental Rights Protection Online*. Edward Elgar Publishing, 2020.

[54] ILGA-Europe. *Hate crime & hate speech*. Accessed = 2021-04-16. 2020. URL: https://www.facebook.com/help/212722115425932.

[55] Chikashi Nobata et al. “Abusive Language Detection in Online User Content”. In: *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Apr. 2016. DOI: 10.1145/2872427.2883062. URL: https://doi.org/10.1145%2F2872427.2883062.

[56] Facebook. *Facebook Hate speech policy*. Accessed = 2021-04-14. 2021. URL: https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech.

[57] YouTube. *Hate speech policy*. Accessed = 2021-04-26. 2021. URL: https://support.google.com/youtube/answer/2801939?hl=en..

[58] Twitter. *Coordinated harmful activity*. Accessed = 2021-04-26. 2021. URL: https://help.twitter.com/en/rules-and-policies/coordinated-harmful-activity.

[59] Tariq Habib Afridi et al. “A Multimodal Memes Classification: A Survey and Open Research Issues”. In: *Innovations in Smart Cities Applications Volume 4*. Springer International Publishing, 2021, pp. 1451–1466. DOI: 10.1007/978-3-030-66840-2_109. URL: https://doi.org/10.1007%2F978-3-030-66840-2_109.

[60] Pradeep K. Atrey et al. “Multimodal fusion for multimedia analysis: a survey”. In: *Multimedia Systems* 16.6 (Apr. 2010), pp. 345–379. DOI: 10.1007/s00530-010-0182-0. URL: https://doi.org/10.1007%2Fs00530-010-0182-0.

[61] Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. “Detecting Sexist MEME On The Web: A Study on Textual and Visual Cues”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE. 2019, pp. 166–170.

[62] Abhinav Sethy and Bhuvana Ramabhadran. “Bag-of-word normalized n-gram models”. In: *Interspeech 2008*. ISCA, Sept. 2008. DOI: 10.21437/interspeech.2008-265. URL: https://doi.org/10.21437%2Finterspeech.2008-265.

[63] M. Hassaballah, Aly Amin Abdelmgeid, and Hammam A. Alshazly. “Image Features Detection, Description and Matching”. In: *Image Feature Detectors and Descriptors*. Springer International Publishing, 2016, pp. 11–45. DOI: 10.1007/978-3-319-28854-3_2. URL: https://doi.org/10.1007%2F978-3-319-28854-3_2.

[64] I. Gallo et al. “Image and Encoded Text Fusion for Multi-Modal Classification”. In: *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, Dec. 2018. DOI: 10.1109/dicta.2018.8615789. URL: https://doi.org/10.1109%2Fdicta.2018.8615789.

[65] Yue Zhang, Qi Liu, and Linfeng Song. “Sentence-State LSTM for Text Representation”. In: (2018). DOI: 10.18653/v1/p18-1030. URL: https://doi.org/10.18653%2Fv1%2Fp18-1030.

[66] Tao Chen et al. “Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN”. In: *Expert Systems with Applications* 72 (Apr. 2017), pp. 221–230. DOI: 10.1016/j.eswa.2016.10.065. URL: https://doi.org/10.1016%2Fj.eswa.2016.10.065.

[67] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[68] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M. Ingersoll. “An Introduction to Logistic Regression Analysis and Reporting”. In: *The Journal of Educational Research* 96.1 (Sept. 2002), pp. 3–14. DOI: 10.1080/00220670209598786. URL: https://doi.org/10.1080%2F00220670209598786.

[69] Andrew McCallum, Kamal Nigam, et al. “A comparison of event models for naive bayes text classification”. In: *AAAI-98 workshop on learning for text categorization*. Vol. 752. 1. Citeseer. 1998, pp. 41–48.

[70] Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i-Nieto. “Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation”. In: *arXiv preprint arXiv:1910.02334* (2019).

[71] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[72] Sayan Goswami. “Reddit Memes Dataset”. In: *Recovered by https://www.kaggle.com/s memes-dataset* (2021).

[73] George-Alexandru Vlad et al. “UPB @ DANKMEMES: Italian Memes Analysis - Employing Visual Models and Graph Convolutional Networks for Meme Identification and Hate Speech Detection”. In: *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*. Accademia University Press, 2020, pp. 288–293.

[74] Liang Yao, Chengsheng Mao, and Yuan Luo. “Graph Convolutional Networks for Text Classification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 7370–7377. DOI: 10.1609/aaai.v33i01.33017370. URL: https://doi.org/10.1609%2Faaai.v33i01.33017370.

[75] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[76] Telmo Pires, Eva Schlinger, and Dan Garrette. “How Multilingual is Multilingual BERT?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/p19-1493. URL: https://doi.org/10.18653%2Fv1%2Fp19-1493.

[77] Alexis Conneau et al. “Unsupervised Cross-Lingual Representation Learning for Speech Recognition”. In: *Interspeech 2021*. ISCA, Aug. 2021. DOI: 10.21437/interspeech.2021-329. URL: https://doi.org/10.21437%2Finterspeech.2021-329.

[78] Riza Velioglu and Jewgeni Rose. “Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge”. In: *arXiv preprint arXiv:2012.12975* (2020).

[79] Chhavi Sharma et al. “SemEval-2020 Task 8: Memotion Analysis–The Visuo-Lingual Metaphor!” In: *arXiv preprint arXiv:2008.03781* (2020).

[80] Piyush Sharma et al. “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2556–2565. DOI: 10.18653/v1/P18-1238. URL: http://dx.doi.org/10.18653/v1/P18-1238.

[81] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. “Multimodal Learning For Hateful Memes Detection”. In: *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, July 2021. DOI: 10.1109/icmew53276.2021.9455994. URL: https://doi.org/10.1109%2Ficmew53276.2021.9455994.

[82] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (June 2017), pp. 1137–1149. DOI: 10.1109/tpami.2016.2577031. URL: https://doi.org/10.1109%2Ftpami.2016.2577031.

[83] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48. URL: https://doi.org/10.1007%2F978-3-319-10602-1_48.

[84] Liunian Harold Li et al. “Visualbert: A simple and performant baseline for vision and language”. In: *arXiv preprint arXiv:1908.03557* (2019). DOI: 10.48550/arXiv.1908.03557. URL: https://doi.org/10.48550/arXiv.1908.03557.

[85] Jiasen Lu et al. “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks”. In: *Advances in neural information processing systems* 32 (2019).

[86] Hao Tan and Mohit Bansal. “LXMERT: Learning Cross-Modality Encoder Representations from Transformers”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Association for Computational Linguistics, 2019. DOI: 10.18653/v1/d19-1514. URL: https://doi.org/10.18653%2Fv1%2Fd19-1514.

[87] Yen-Chun Chen et al. "UNITER: UNiversal Image-TExt Representation Learning". In: *Computer Vision ECCV 2020*. Springer International Publishing, 2020, pp. 104–120. DOI: 10.48550/arXiv.1909.11740. URL: https://doi.org/10.48550/arXiv.1909.11740.

[88] Xiujun Li et al. "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks". In: *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 121–137. DOI: 10.1007/978-3-030-58577-8_8. URL: https://doi.org/10.1007%2F978-3-030-58577-8_8.

[89] Joseph Redmon and Ali Farhadi. "YOLO9000: Better, Faster, Stronger". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. DOI: 10.1109/cvpr.2017.690. URL: https://doi.org/10.1109%2Fcvpr.2017.690.

[90] Oriol Vinyals et al. "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (Apr. 2017), pp. 652–663. DOI: 10.1109/tpami.2016.2587640. URL: https://doi.org/10.1109%2Ftpami.2016.2587640.

[91] Vung Pham, Chau Pham, and Tommy Dang. "Road Damage Detection and Classification with Detectron2 and Faster R-CNN". In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2020. DOI: 10.1109/bigdata50022.2020.9378027. URL: https://doi.org/10.1109%2Fbigdata50022.2020.9378027.

[92] Fei Yu et al. ""Ernie-vil: Knowledge enhanced vision-language representations through scene graph"". In: *arXiv preprint arXiv:2006.16934* 1 (2020), p. 12. DOI: 10.48550/arXiv.2006.16934. URL: https://doi.org/10.48550/arXiv.2006.16934.

[93] Weibo Zhang et al. "Hateful Memes Detection via Complementary Visual and Linguistic Networks". In: *arXiv preprint arXiv:2012.04977* (2020).

[94] Bhargav Srinivasa-Desikan. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd, 2018.

[95] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[96] Ron Zhu. "Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution". In: *arXiv preprint arXiv:2012.08290* (2020).

[97] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation". In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2021. DOI: 10.1109/wacv48630.2021.00159. URL: https://doi.org/10.1109%2Fwacv48630.2021.00159.

[98] Raul Gomez et al. "Exploring Hate Speech Detection in Multimodal Publications". In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Mar. 2020. DOI: 10.1109/wacv45572.2020.9093414. URL: https://doi.org/10.1109%2Fwacv45572.2020.9093414.

[99] Anh Dang et al. "An offline–online visual framework for clustering memes in social media". In: *From Social Data Mining and Analysis to Prediction and Community Detection*. Springer, 2017, pp. 1–29. DOI: 10.1007/978-3-319-51367-61. URL: https://doi.org/10.1007/978-3-319-51367-61.

[100] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[101] Alexander M Robertson and Peter Willett. "Applications of n-grams in textual information systems". In: *Journal of Documentation* (1998).

[102] Nikolaos Aletras and Mark Stevenson. "Measuring the similarity between automatically generated topics". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2:*

Short Papers. 2014, pp. 22–27. DOI: 10.3115/v1/E14-4005. URL: https://doi.org/10.3115/v1/E14-4005.

[103] Yuxin Wu et al. *Detectron2*. https://github.com/facebookresearch/detectron2. 2019.

[104] Chunliang Lu et al. "Web entity detection for semi-structured text data records with unlabeled data". In: *International Journal of Computational Linguistics and Applications* 4.2 (2013), pp. 135–150.

[105] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164. DOI: 10.48550/arXiv.1411.4555. URL: https://doi.org/10.48550/arXiv.1411.4555.

[106] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[107] Lisa Bonheme and Marek Grzes. "SESAM at SemEval-2020 task 8: investigating the relationship between image and text in sentiment analysis of memes". In: (2020). DOI: 10.18653/v1/2020.semeval-1.102. URL: https://doi.org/10.18653/v1/2020.semeval-1.102.

[108] Edward J Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *arXiv preprint arXiv:2106.09685* (2021). URL: https://arxiv.org/abs/2106.09685.

[109] Dor Bank, Noam Koenigstein, and Raja Giryes. "Autoencoders". In: *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (2023), pp. 353–374. DOI: 10.48550/arXiv.2003.05991. URL: https://doi.org/10.48550/arXiv.2003.05991.

[110] Shraman Pramanick et al. "MOMENTA: A multimodal framework for detecting harmful memes and their targets". In: *arXiv preprint arXiv:2109.05184* (2021).

[111] Shraman Pramanick et al. “Detecting harmful memes and their targets”. In: *arXiv preprint arXiv:2110.00413* (2021). DOI: 10.48550/arXiv.2110.00413. URL: https://doi.org/10.48550/arXiv.2110.00413.

[112] Muhammad Mateen et al. “Fundus image classification using VGG-19 architecture with PCA and SVD”. In: *Symmetry* 11.1 (2018), p. 1. DOI: 10.3390/sym11010001. URL: https://doi.org/10.3390/sym11010001.

[113] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *arXiv preprint arXiv:1910.01108* (2019). DOI: 10.48550/arXiv.1910.01108. URL: https://doi.org/10.48550/arXiv.1910.01108.

[114] Di Qi et al. “Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data”. In: *arXiv preprint arXiv:2001.07966* (2020).

[115] Tam Sakirin and Siddartha Kusuma. “A Survey of Generative Artificial Intelligence Techniques”. In: *Babylonian Journal of Artificial Intelligence* 2023 (2023), pp. 10–14.

[116] Xinlei Chen et al. “Microsoft coco captions: Data collection and evaluation server”. In: *arXiv preprint arXiv:1504.00325* (2015). DOI: 10.48550/arXiv.1504.00325. URL: https://doi.org/10.48550/arXiv.1504.00325.

[117] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015), pp. 91–99. DOI: 10.48550/arXiv.1506.01497. URL: https://doi.org/10.48550/arXiv.1506.01497.

[118] Shivam Sharma et al. “Detecting and understanding harmful memes: A survey”. In: *arXiv preprint arXiv:2205.04274* (2022). DOI: 10.48550/arXiv.2205.04274. URL: https://doi.org/10.48550/arXiv.2205.04274.

[119] Roy Ka-Wei Lee et al. “Disentangling hate in online memes”. In: *Proceedings of the 29th ACM international conference on multimedia*. 2021, pp. 5138–5147. DOI: 10.1145/3474085.3475625. URL: https://doi.org/10.1145/3474085.3475625.

[120] Hannah Rose Kirk et al. “Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset”. In: *arXiv preprint arXiv:2107.04313* (2021). DOI: 10.48550/arXiv.2107.04313. URL: https://doi.org/10.48550/arXiv.2107.04313.

[121] Tae Kyun Kim. “T test as a parametric statistic”. In: *Korean journal of anesthesiology* 68.6 (2015), p. 540.

[122] Gitanjali Kumari, Anubhav Sinha, and Asif Ekbal. “Unintended Bias Detection and Mitigation in Misogynous Memes”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 2719–2733.

[123] Jiayang Wu et al. “Multimodal large language models: A survey”. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE. 2023, pp. 2247–2256.

[124] Ggaliwango Marvin et al. “Prompt Engineering in Large Language Models”. In: *International Conference on Data Intelligence and Cognitive Informatics*. Springer. 2023, pp. 387–402.

[125] JiaLu Xing et al. “A survey of efficient fine-tuning methods for Vision-Language Models—Prompt and Adapter”. In: *Computers & Graphics* (2024).

[126] Douwe Kiela et al. “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes”. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2020, pp. 2611–2624.

[127] Shardul Suryawanshi et al. “Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text”. In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. 2020, pp. 32–41.

[128] Rui Cao et al. “Prompting for Multimodal Hateful Meme Classification”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 321–332.

[129] Shraman Pramanick et al. “Detecting Harmful Memes and Their Targets”. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021, pp. 2783–2796.

[130] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv preprint* (2019). eprint: 1907.11692.

[131] Giulia Rizzi et al. “Recognizing misogynous memes: Biased models and tricky archetypes”. In: *Information Processing & Management* 60.5 (2023), p. 103474.

[132] Fan Wu et al. “Fuser: An enhanced multimodal fusion framework with congruent reinforced perceptron for hateful memes detection”. In: *Information Processing & Management* 61.4 (2024), p. 103772.

[133] Shawly Ahsan et al. “A Multimodal Framework to Detect Target Aware Aggression in Memes”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 2487–2500.

[134] Zekun Li et al. “Guiding large language models via directional stimulus prompting”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2024.

[135] Cole E Short and Jeremy C Short. “The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation”. In: *Journal of Business Venturing Insights* 19 (2023), e00388.