Information Fusion 106 (2024) 102269

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

KERMIT: Knowledge-EmpoweRed Model In harmful meme deTection

Biagio Grasso, Valerio La Gatta *, Vincenzo Moscato, Giancarlo Sperlì

Department of Electrical Engineering and Information Technology (DIETI), University of Naples Federico II, Via Claudio 21, Naples, Italy

ARTICLE INFO

Keywords:
Internet memes
Harmful meme detection
Knowledge-informed decision-making
Multimodal disinformation mining

ABSTRACT

Internet memes, while often humorous in nature, can be used to spread hate speech, toxic content, and disinformation across the digital information ecosystem. As a result, detecting harmful memes has become a crucial task for maintaining online safety and fostering responsible online behavior. Prior research in this field has mainly targeted multimodal internal aspects of memes, specifically the image and text modalities, and has sought to interpret their significance by analyzing intra- and inter-modality signals via sophisticated visual-language models. However, understanding the message of a (harmful) meme entails tacit background knowledge, which is not explicitly expressed in the meme itself, but rather relies on cultural references, shared knowledge, and social context. In this paper, we propose KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), a novel framework which incorporates and uses external knowledge into the process of identifying harmful memes. Specifically, KERMIT builds the meme's *knowledge-enriched information network* by integrating internal entities of the meme with relevant external knowledge obtained from ConceptNet. Subsequently, the framework employs a dynamic learning mechanism that leverages memory-augmented neural networks and attention mechanisms to discern the most informative knowledge for accurate classification of harmful memes. Our experiments on four benchmark datasets demonstrate that KERMIT effectively utilizes external knowledge to improve classification performance compared to several state-of-the-art baselines. Overall, the findings of this study shed light on the complex nature of Internet memes and highlight the importance of knowledge-informed decision-making for harmful meme detection.

1. Introduction

The utilization of internet memes, which are often recognized for their humorous nature, has extended beyond their original purpose to encompass the dissemination of hate speech, toxic content, and disinformation on social media platforms. An illustrative example is the "Pepe The Frog"[1] meme, which has now become a symbol closely associated with far-right and white supremacist groups [1]. Similarly, the share of the former US President Donald Trump of a meme featuring two QAnon slogans is widely regarded as one of his most overt acknowledgments of the QAnon conspiracy theory.[2] As a result, detecting harmful memes accurately and efficiently has become a crucial task for maintaining online safety and fostering responsible online behavior.

In recent years, this issue has drawn attention from scholars and practitioners in diverse research fields, including natural language processing [2,3], computer vision [4,5], and social media analysis [6]. In particular, detecting harmful memes presents unique challenges compared to other forms of hate speech detection, as memes often employ visual and textual elements in combination to convey their messages. For this reason, previous work has predominantly examined the internal content of memes by employing multimodal representation learning strategies to analyze intra- and inter-modality signals between image and text modalities [7,8]. Specifically, the state-of-the-art architectures leverage pre-trained visual-language models (e.g., VisualBERT [9], CLIP [10], MMBT [11]) as they harness the combined information from both text and images to capture the nuanced meanings of a meme.

However, (harmful) memes also rely on cultural references, shared knowledge, and social context to convey their intended meaning [12,13]. In other words, understanding the message of a meme entails tacit background knowledge, which is not explicitly expressed in the meme itself, but rather relies on the viewer's familiarity with certain contextualized aspects of the world. For instance, the meme depicted in Fig. 1(a) initially appears innocuous, showing a beautiful woman with accompanying text. However, upon further analysis and contextualization with real-world knowledge, it becomes evident that the meme

* Corresponding author.
E-mail addresses: bia.grasso@studenti.unina.it (B. Grasso), valerio.lagatta@unina.it (V. La Gatta), vincenzo.moscato@unina.it (V. Moscato), giancarlo.sperli@unina.it (G. Sperlì).

[1] https://wikipedia/Pepe_The_Frog.jpg

[2] https://cnn.com/qanon-fans-donald-trump.html

https://doi.org/10.1016/j.inffus.2024.102269
Received 18 May 2023; Received in revised form 22 January 2024; Accepted 23 January 2024
Available online 25 January 2024

Fig. 1. Examples of tacit background knowledge exploited by Internet memes.

is making a controversial and provocative statement about beauty and race by suggesting that being white is inherently superior or desirable in terms of physical attractiveness, ascribing the woman's beauty to her race. In a more abstract vein, the meme in Fig. 1(b) juxtaposes a deserted landscape with a phrase typically indicative of affection and popularity. This meme's impact lies in the stark contrast between the desolation of the desert, symbolizing solitude, and the phrase "look how many people", which connotes the presence of a crowd. The viewer's understanding of this dichotomy is essential to grasp the meme's ironic or mocking tone, as it plays on the societal expectations of affection and the reality of isolation. In general, the necessary background knowledge for interpreting memes varies based on their level of abstraction. In the case of the meme in Fig. 1(a), understanding may hinge on recognizing explicit demographic attributes such as race and gender. In contrast, the meme in Fig. 1(b) is more abstract and relies more on an understanding of nuanced concepts related to symbolic or metaphorical imagery.

Prior research in meme analysis has predominantly focused on the basic augmentation of meme text with demographic attributes derived from associated images [8,14,15], often neglecting the structured modeling and integration of contextual knowledge. The general assumption is that large multimodal models intrinsically capture essential background knowledge during their pre-training process [16–18]. However, this approach is inherently limited, as pre-training datasets may not encompass the full spectrum of context-specific information critical for decoding the nuanced and intricate aspects of meme content. Additionally, these models' complexity and opacity impede the clear identification and quantification of biases, posing challenges to their real-world applications [19]. Indeed, while these prejudices could inadvertently aid in detecting harmful memes, a blind reliance on their (biased) pre-trained knowledge results in systems that perpetuate their biases against specific ethnic groups and women [20,21], ultimately leading to unfair censorship of marginalized groups [22,23]. These issues underscore the necessity for more balanced and inclusive approaches in harmful meme detection. Therefore, the direct incorporation of external knowledge into the classification process emerges as a promising strategy to not only enhance the effectiveness of harmful meme detection but also to amplify the real-world applicability of these solutions.

Grounded in the identified need for a more nuanced approach to meme analysis, this paper introduces a novel methodology, KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), designed for the explicit modeling and integration of memes' background knowledge into the detection of harmful memes. KERMIT operates through a two-step framework: Firstly, it builds the meme's *knowledge-enriched information network* by integrating internal entities of the meme with relevant external knowledge obtained from ConceptNet [24]. Subsequently, KERMIT employs a dynamic learning mechanism that leverages memory-augmented neural networks and attention mechanisms to discern the most informative segment of the *knowledge-enriched information network* for accurate classification of harmful memes.

To the best of our knowledge, our study represents the first comprehensive attempt to model meme-related knowledge, encompassing both the meme's internal entities and their relationships, as well as external related knowledge obtained from ConceptNet [24]. Notably, while previous work has explored incorporating unstructured knowledge into the decision-making process, such as web entities detected in image modality [7] or semantic entities retrieved from ConceptNet [25], our proposed framework is the first end-to-end solution that dynamically learns the most relevant knowledge for the task of hateful classification.

Our evaluation encompasses a comprehensive analysis across four benchmark datasets: Facebook Hateful Memes [4], Multimedia Misogyny Dataset (MAMI) [5], MultiOFF [26], and Memotion7k [27], each targeting different aspects and categories of harmful memes.

Our results demonstrate that KERMIT is capable of retrieving relevant contextual knowledge from ConceptNet, and effectively utilizing it in the classification process to enhance performance. Specifically, the proposed system achieves state-of-the-art performance on the Facebook Hateful Memes dataset and performs comparably with the most recent competitors on all other datasets. Overall, this work demonstrates the effectiveness of incorporating external knowledge into the classification process and sets a path for future research in the harmful meme detection, highlighting the significant role that artificial intelligence and knowledge discovery can play in improving content moderation.

The paper is organized as follows: Section 2 provides an overview of prior research on harmful meme detection and memory-augmented neural networks. Section 3 introduces KERMIT, detailing its architecture and functionality. In Section 4, we conduct extensive evaluations of our framework, comparing it to various state-of-the-art baselines. Lastly, Section 5 presents concluding remarks and discusses potential avenues for future research aimed at further enhancing KERMIT's capabilities.

2. Related works

2.1. Harmful meme detection

The increasing employment of memes for spreading harmful content has emerged as a significant concern, prompting an intensification of research efforts in this domain [8]. This heightened focus has advanced harmful meme detection and stimulated research into the specific persuasive techniques in meme messaging [3] and the identification of targets in these memes, such as public figures or political entities [28–30]. Despite the specific task, addressing harmful meme detection entails navigating the unique challenges posed by their multimodal nature, where text and imagery intertwine to deliver complex messages. This aspect distinguishes harmful memes from other multimodal tasks like cross-modal retrieval [31] or visual question-answering [32], where text and images typically convey congruent messages.

In this context, traditional unimodal approaches, focusing on either text or imagery, fall short in capturing the complete context of

harmful memes [4]. Consequently, the predominant approach involves fine-tuning pre-trained visual-language models (e.g., VisualBERT [9], CLIP [10], MMBT [11]), which integrate textual and visual information to understand the meanings and nuances in memes. The effectiveness of such models was demonstrated in the *Facebook Hateful Meme Detection* challenge [33], where its competitors also highlighted how models' ensemble and the source domain of pre-training datasets influence detection performance [34,35].

Furthermore, recent advancements in meme research have focused on the bimodal interplay between text and image modalities. For instance, Hate-CLIPper [36] learns correlations between embeddings of these modalities. Memefier [15] and MHA-Meme [37] assess alignment between text and images at a different granular levels, while DisMultiHate [14] focuses on disentangling visual and textual representations for identifying various hate categories (e.g., religion, race).

The above-mentioned works have centered primarily on the visual and textual components of memes, often neglecting the importance of external knowledge such as cultural references and social context [12,13]. Typically, memes' context is represented using additional unstructured data like image tags [7,29,30] or demographic features [14,15]. Evidently, this solution proves inadequate for more abstract memes (see Fig. 1(b)) that lack straightforward demographic attributes or web entities. In most of the cases, the assumption is that vision-language models inherently acquire necessary background knowledge during pre-training. This perspective is challenged by methodologies like PromptHate [38] and Pro-cap [39], which directly instruct a language model with a prompt containing the meme's text and caption. However, this blind reliance on pre-trained language models has its limitations, since pre-training datasets may not cover context-specific information essential for interpreting complex meme content. Moreover, the complexity and opacity of large language models hinder full understanding of their biases [19], and can result in systems that perpetuate these biases [22,23] and unfairly censor marginalized groups, as shown for harmful meme detection systems based on pre-trained models [21].

For these reasons, we approach the problem from a different, yet complementary, perspective by proposing the explicit (graph) modeling and integration of human common-sense knowledge into the harmful classification process. It is worth noting that our methodology, KERMIT, does not seek to replace any vision-language model, but rather aims to unveil or even exploit their biases in order to achieve a more informed decision making process. To the best of our knowledge, KnowMeme [25] is the most relevant research endeavor to achieve knowledge-informed hateful meme classification. That is, the system builds a graph representing the meme content and related knowledge, retrieved from ConceptNet, and then performs graph classification to detect whether the meme is harmful or not.

However, our framework KERMIT differs from KnowMeme in several key points. First, KnowMeme builds the meme graph without considering relationships, whereas KERMIT incorporates a modeling approach that takes into account the relational information within the meme graph. Second, KERMIT employs a dynamic learning mechanism to determine the most informative portion of the graph for meme classification. Finally, in contrast to KnowMeme, which solely relies on the meme graph for the hateful classification, KERMIT adopts a hybrid approach that combines explicit knowledge from the meme graph and a pre-trained vision-language model.

2.2. Memory-augmented neural networks

In the context of informed decision-making, memory-augmented neural networks [40] (MANNs) integrate the capabilities of neural networks to discern intricate patterns in data with the ability to store and retrieve information from an external memory. These models have demonstrated their effectiveness in a wide range of tasks, including question answering [41], image captioning [42], and video analysis [43]. Specifically, MANNs extend traditional neural networks with an external *memory block* that stores task-related knowledge, such as contextual information, factual data, or domain expertise. Formally, the *memory block* consists of several *memory slots*, each containing an elementary *bucket* of knowledge. The representation of each *bucket* is dependent on the specific task and is a design choice, with possible options including unstructured text [44], tables [45] or graphs [46].

In the context of harmful meme detection, incorporating external knowledge can be beneficial for better understanding the intent, the tone, and the reliability of memes as well as catching its cultural allusions and social context [25]. For this reason, our framework KERMIT leverages a MANN to store the *knowledge-enriched information network* representing the meme's entities and their related common-sense knowledge retrieved from ConceptNet. Specifically, the *memory block* of KERMIT includes several *buckets*, each one storing a part of the *knowledge-enriched information network*. In addition, KERMIT also leverages an attention mechanism to dynamically learn the most informative *bucket(s)* for accurate classification of harmful memes. This approach allows our model to integrate external knowledge into the decision-making process, improving its ability to detect harmful content in memes by considering contextual information and relevant knowledge from external sources.

3. Methodology

In this section, we formalize the problem of knowledge-informed harmful meme detection and describe the architecture of KERMIT highlighting any detail about how we extract meme's related knowledge as well as how we incorporate such knowledge within the classification process.

3.1. Problem formulation

We consider an internet meme $\mathcal{M}$ as a complex information piece composed of an image $\mathcal{I}_M$ and an embedded text $\mathcal{T}_M$. In particular, we define the set of meme's entities as $S_M = \{s_1, s_2, \dots, s_h\}$, s_i being a meaningful concept within $\mathcal{I}_M$ or $\mathcal{T}_M$ (e.g., a face within the image, an entity name within the embedded text). In addition, we also consider the set of relationships between these entities as $\mathcal{P}_M = \{p_{ij}\} = \{(s_i, s_j) | s_i, s_j \in S_M\}$. For instance, for the meme in Fig. 1(a), the possible relationships include the woman smiling at the camera and her beauty being attributed to her white parents. As a result, we define the *meme graph* $\mathcal{G}_M = \{S_M, \mathcal{P}_M\}$ which represents the internal knowledge expressed directly within $\mathcal{M}$.

Building upon the premise that understanding a meme requires background knowledge, we integrate an external knowledge base $\mathcal{K}$, and define the meme's background knowledge $\mathcal{K}_M$ as the subset of facts within $\mathcal{K}$ pertinent to any entity in $\mathcal{M}$. For instance, in the meme from Fig. 1(a), relevant background knowledge for the textual modality might include the connotations of "white" as a race descriptor and its associations with racism, as well as the interpretation of "pretty" as a synonym of "dolly" and "frivolous", implying a beautiful yet superficial woman. Consequently, we define $\widetilde{\mathcal{G}}_M$ as the meme's *knowledge-enriched information network*, which is essentially the meme graph $\mathcal{G}_M$ augmented with the external knowledge $\mathcal{K}_M$.

As a result, we frame the task of knowledge-informed harmful meme detection as a multimodal classification problem. This involves to learn a binary decision function $f : \mathcal{M} = \{\mathcal{I}_M, \mathcal{T}_M, \widetilde{\mathcal{G}}_M\} \rightarrow y \in \{0, 1\}$. Notably, this formulation does not depend on the strategy used to construct $\widetilde{\mathcal{G}}_M$ as well as it could be seamlessly generalized to handle multi-class scenarios.

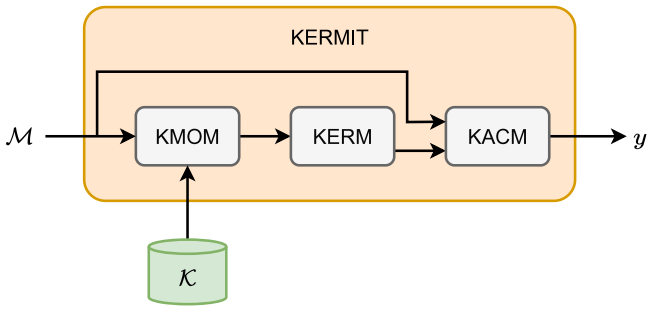

Fig. 2. The high-level architecture of KERMIT: The KMOM module leverages the input meme $\mathcal{M}$ and an external knowledge base $\mathcal{K}$ and generates the meme's *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$. Subsequently, the KERM module embeds $\widetilde{\mathcal{G}}_M$ into a lower-dimensional space. Finally, the KACM module integrates the knowledge stored in $\widetilde{\mathcal{G}}_M$ together with the input meme for hateful classification.

3.2. Our framework

Fig. 2 depicts the architecture of the proposed framework. Specifically, KERMIT consists of three main modules: the Knowledge Modeling and Organization Module (KMOM), the Knowledge Embedding Representation Module (KERM), and the Knowledge-Augmented Classification Module (KACM).

The KMOM module is responsible for generating the meme's *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$, which merges meme's entities and their related background knowledge retrieved from $\mathcal{K}$. Subsequently, the KERM module is responsible for embedding $\widetilde{\mathcal{G}}_M$ in a lower-dimensional, continuous latent space that maintains the topological and semantic information of its nodes and edges. Finally, the KACM module integrates information from the input meme $\mathcal{M}$ and its *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$ to perform the hateful classification.

The following sections provide a detailed description of the architecture of each module.

3.2.1. Knowledge modeling and organization (KMOM)

This module consists of a two-stage process for generating the meme's *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$. In the first stage, we construct the *meme graph* $\mathcal{G}_M$ from the meme content. Subsequently, we generate $\widetilde{\mathcal{G}}_M$ by enriching $\mathcal{G}_M$ with additional knowledge retrieved from ConceptNet [24].

Meme graph. The internal concepts of a meme are shaped by the entities and relationships present within its multimodal content. Fig. 3 illustrates the process of extracting these concepts and building $\mathcal{G}_M$. Firstly, the meme's embedded text and caption are retrieved using OCR techniques and the BLIP vision-language model [47], respectively. Subsequently, in line with previous work that focuses on constructing knowledge graphs from unstructured text [48–50], the *Graph Extraction* block performs Part-Of-Speech (POS) tagging to identify the set of nodes $\mathcal{S}_M$ in $\mathcal{G}_M$ from both the caption and the embedded text. For instance, for the meme in Fig. 3, we extract the nouns *woman*, *camera* (resp., *parents*, *pretty*) from the caption (resp., the embedded text) of the meme.

Next, we leverage dependency parsing to extract the set of relationships $\mathcal{P}_M$ between the above-mentioned nodes. In particular, given the input text — either the meme's caption or embedded text — the parsing tree provides the relationships between the words in the input based on the syntactic structure of the input sentence (e.g., subject-verb or adjective-noun relationships). Since we are interested in the general relationships between the words rather than their specific grammatical dependencies, we replace the typologies of the dependence (such as subject modifier or coordination conjunction) with the general relation *relatedTo*. This simplifies the representation of the relationships in the *meme graph* while preserving their dependencies.

Finally, we merge the dependency trees of the caption and embedded text by connecting their root nodes and common words (e.g., the edges connecting the terms "pretty" or the terms "am" and "smiling" in Fig. 3). This process results in the final meme graph $\mathcal{G}_M$, which is not formally a dependency tree, but rather a representation of the soft-connection between the caption and the embedded text.

Overall, by leveraging dependency parsing and simplifying the relationships to a general *relatedTo* category, we conjecture that $\mathcal{G}_M$ can effectively provide a more comprehensive understanding of the concepts represented within the meme content.

Algorithm 1 $\widetilde{\mathcal{G}}_M$ Generation

1: **procedure** KnowledgeEnrichment($\mathcal{G}_M$, depth l)
2: initialize $\widetilde{\mathcal{G}}_M = \mathcal{G}_M$
3: **for** each node $s_i \in \mathcal{G}_M$ **do**
4: **if** $pos\_tag(s_i) \in \{noun, adj, verb\}$ **then**
5: $k\_entities, k\_relationships \leftarrow GetConceptNet(s_i, l)$
6: add $k\_entities$ to $\widetilde{\mathcal{G}}_M$
7: add $k\_relationships$ to $\widetilde{\mathcal{G}}_M$
8: **return** $\widetilde{\mathcal{G}}_M$

Knowledge enrichment. In order to enhance the information contained in the meme graph, we leverage ConceptNet [24] as the external knowledge base to construct the meme's *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$.

We choose to leverage ConceptNet over other factual knowledge bases, such as WikiData [51], due to our intuition that the common-sense knowledge stored in ConceptNet might be more helpful in capturing the true meaning of the meme. For instance, the word "black" may be factually related to the term "color" in WikiData, but when considering common-sense knowledge, "black" can also be intended as an offensive word related to the words "negro" and "racist". This nuanced understanding is important when analyzing memes, as their meaning is often heavily dependent on context and connotation. We will validate this choice in the experiments.

Fig. 4 shows an example of a subset of knowledge extracted from the embedded text of the meme in Fig. 3 and also highlights the recursive nature of the knowledge retrieval algorithm. For example, we query ConceptNet to retrieve concepts related to *pretty*, obtaining that *pretty* is related to *putty* and *dolly*, is a synonym of *lovely* and *flower* has the property of being *pretty*. In addition, the term *dolly* is recursively used to retrieve additional information (i.e., *truck*, *frivolous*). The recursion depth is a parameter of the knowledge extraction process and controls the number of nested queries performed to ConceptNet; the larger the *recursion depth* is, the greater is the amount of external knowledge incorporated into the final $\widetilde{\mathcal{G}}_M$.

Specifically, Algorithm 1 outlines the methodology of the knowledge enrichment process. The algorithm takes as input the initial meme graph $\mathcal{G}_M$ and the recursion depth l. Concretely, the algorithm starts with the initialization of $\widetilde{\mathcal{G}}_M$ to be identical to $\mathcal{G}_M$ (line 2). Subsequently, for each node s_i in $\mathcal{G}_M$ whose pos tag is noun, verb or adjective (lines 3–4), we query ConceptNet to retrieve a set of facts related to s_i (line 5). Specifically, $k\_entities$ and $k\_relationships$ represent the nodes and the edges that need to be iteratively added to $\widetilde{\mathcal{G}}_M$ (lines 6–7). For example, the green region in Fig. 4 contains a subset of nodes and edges that enrich the caption of the meme in Fig. 3.

Furthermore, the *GetConceptNet* function implements the recursive querying mechanism, as described in Algorithm 2. This procedure takes in input a general entity e to be queried for and the recursion depth l. For instance, Fig. 4 shows two levels of knowledge ($l = 2$) retrieved from the embedded text of the meme in Fig. 3. In particular, Algorithm

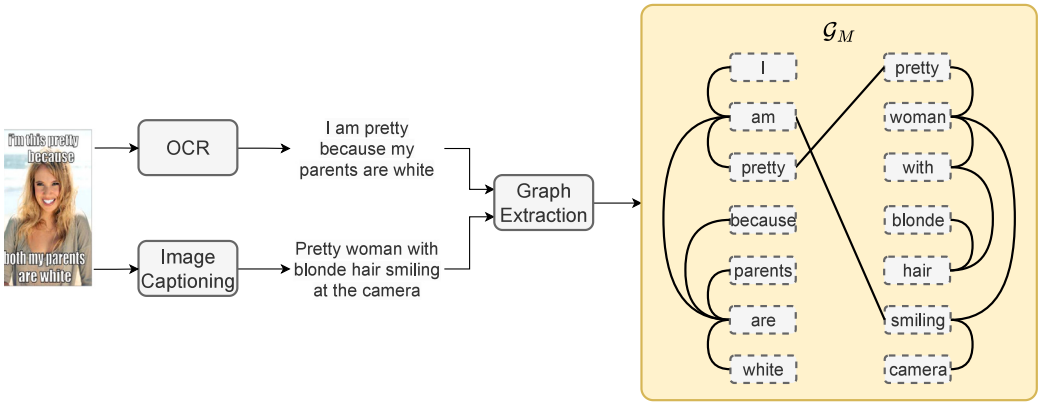

Fig. 3. The workflow to build the meme graph $\mathcal{G}_M$.

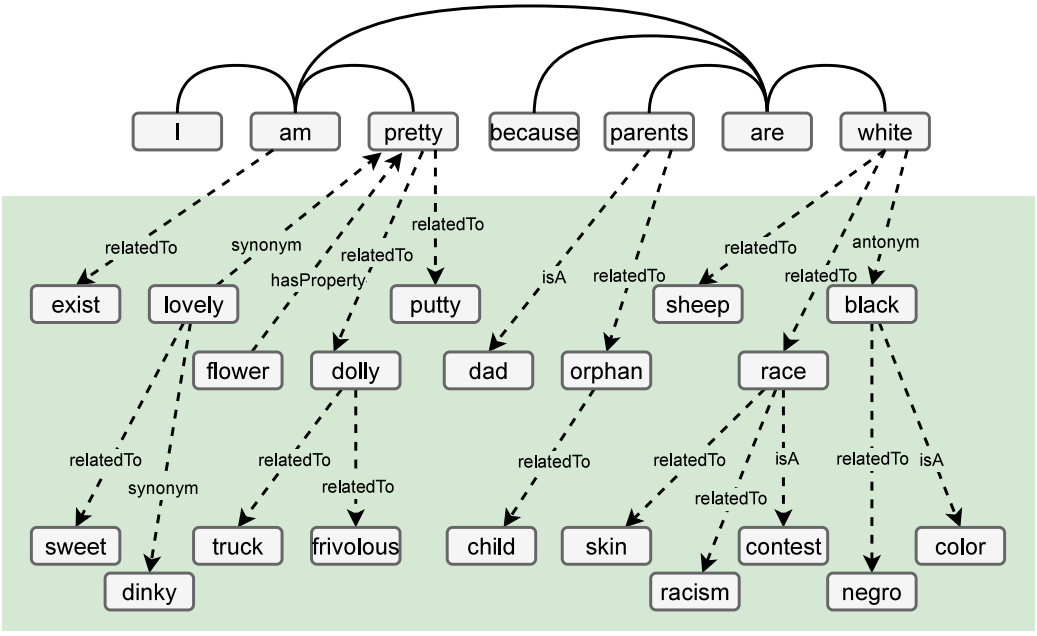

Fig. 4. Knowledge enrichment: common-sense knowledge retrieved from ConceptNet related to the embedded text of the meme in Fig. 3.

Algorithm 2 Querying ConceptNet

```
1:  procedure GetConceptNet(entity e, depth l)
2:      initialize S_K as empty entities' set
3:      initialize E_K as empty relationships' set
4:      if l = 0 then
5:          return S_K, E_K
6:      triples ← query_conceptnet_api(e)
7:      for each triple = (s, p, o) ∈ triples do
8:          S_K ← o ∪ S_K
9:          S_K ← s ∪ S_K
10:         E_K ← (s, p, o) ∪ E_K
11:         if l > 1 then
12:             sub_S_K, sub_E_K ← GetConceptNet(o, l − 1)
13:             sub_S_K, sub_E_K ← GetConceptNet(s, l − 1)
14:             S_K ← sub_S_K ∪ S_K
15:             E_K ← sub_E_K ∪ E_K
16:     return S_K, E_K
```

2 initially sets the *entities* $\mathcal{S}_K$ and *relationships* $\mathcal{E}_K$ objects to empty sets (lines 2–3). Subsequently, if the recursion depth l has not been reached yet (lines 4–6), we retrieve the set of *triples* related to e (line 7). In particular, each triple consists of a subject s and an object o, related between each other through the predicate p (e.g., the triple *white, relatedTo, race* in Fig. 4). Subsequently, the algorithm recursively builds sub-graphs for s and o in each triple (lines 8–16). The (sub-)entities $sub\_\mathcal{S}_K$ and (sub-)relationships $sub\_\mathcal{E}_K$ returned from each recursive call are added to the $\mathcal{S}_K$ and $\mathcal{E}_K$ sets, respectively. Importantly, the meaning of the retrieved predicates is contingent on the design of ConceptNet and encompasses diverse types of relationships (e.g., *relatedTo*, *isA*, *isAntonym*).

Looking at Fig. 4, it can be noticed that the added knowledge includes useful entities, such as *dolly*, *racism* and *negro*, which could help to classify the meme as hateful or not. However, the presence of polysemic words such as *race* and *dolly* leads to unrelated terms such as *truck* and *contest*. Overall, the final knowledge-enriched information network $\widetilde{\mathcal{G}}_M$ is a semantic attributed graph whose nodes include potentially helpful concepts to understand the meme and whose edges represent how those concepts are related to each other.

3.2.2. Knowledge embedding representation (KERM)

Once having retrieved the meme's *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$, we aim at generating robust vector representations for the nodes (i.e., words) within $\widetilde{\mathcal{G}}_M$. . These embeddings should encapsulate

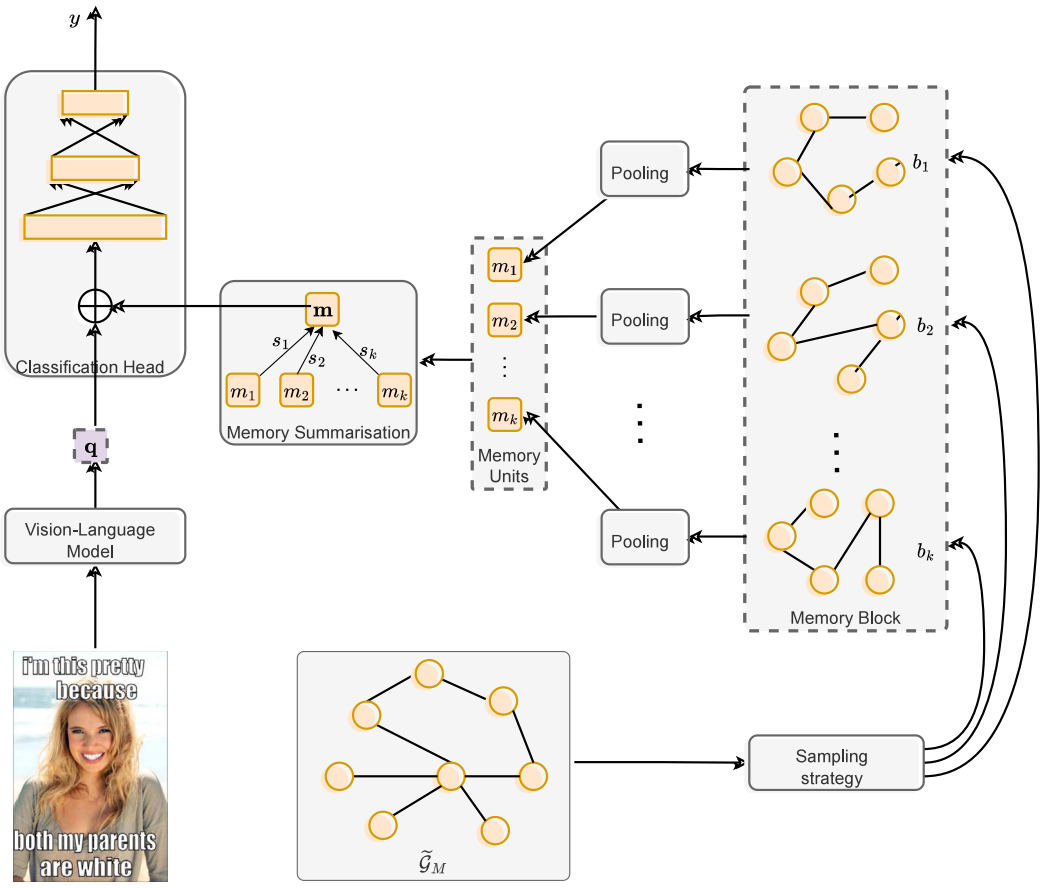

Fig. 5. The architecture of the KACM module: the *vision-language model* encodes the input meme into a vector representation **q**. The *sampling* and *pooling* components provide the memory buckets b_i and their vector representations m_i, respectively. Next, the *memory summarization* component dynamically learns the most informative knowledge context **m** for the hateful classification. Finally, the *classification head* performs the hateful classification based on the merged representations of the meme and summarized memory.

both the semantic and structural information associated with each node. For this reason, we consider $\widetilde{\mathcal{G}}_M$ as a heterogeneous graph with two node types: (i) nodes extracted from the meme's caption and embedded text, i.e., those belonging to the original meme graph; (ii) nodes retrieved from ConceptNet; i.e., those added during the knowledge enrichment process.

Our embedding process involves a two-step methodology: (1) firstly, for each word $w \in \widetilde{\mathcal{G}}_M$, BERT [52] is utilized to produce an initial semantic embedding. Specifically, we input w into BERT and extract the embedding $x_w \in \mathcal{R}^n$ of the [CLS] token, n being the latent dimension. (2) Next, we utilize HIN2Vec [53] to infuse these embeddings with structural information pertinent to $\widetilde{\mathcal{G}}_M$. HIN2Vec processes the entire $\widetilde{\mathcal{G}}_M$, with each node characterized by its BERT-generated embedding. The algorithm unsupervisedly refines the embeddings to incorporate local connectivity patterns within $\widetilde{\mathcal{G}}_M$, thereby augmenting the nodes' embeddings with both semantic and structural information.

Despite our specific choice, we emphasize that KERMIT's design is parametric with respect to any embedding procedure. We will critically assess the effects of our choices in the experimental section of our study. In addition, it is worth to note that the decision to perform node embedding rather than whole graph embedding depends on knowledge heterogeneity in $\widetilde{\mathcal{G}}_M$. Indeed, when embedding the whole graph into a single low-dimensional vector, we lose the ability to distinguish between valuable and extraneous information. By contrast, we extract individual embeddings for each node and defer the identification of useful information to the next stage of the framework, i.e., the *Knowledge-Augmented Classification Module*.

3.2.3. Knowledge-augmented classification (KACM)

Our focus now shifts to how we inject the knowledge-enriched information network into the classification process. Specifically, the *Knowledge-Augmented Classification Module* (KACM) is responsible for performing knowledge-informed hateful classification, and its architecture, depicted in Fig. 5, is inspired by memory-augmented neural networks [40,54]. The system is designed to take advantage of the rich knowledge stored in $\widetilde{\mathcal{G}}_M$ by incorporating it into a *memory block* that can be accessed during classification. Concretely, this memory element comprises several slots, namely *buckets*, each one storing a subset of knowledge stored in $\widetilde{\mathcal{G}}_M$. Given the input meme and the memory component, the workflow is as follows:

(i) The Vision-Language model encodes the input meme into a vector representation $q \in \mathcal{R}^N$;
(ii) The pooling module independently encodes each bucket b_i into vector representations $m_i \in \mathcal{R}^M$;
(iii) The memory summarization module dynamically learns the most relevant context for the hateful classification task and fuses the memory units into a final vector representation.
(iv) The classification head performs the final hateful classification based on the merged representations of the meme q and (summarized) memory **m**.

The following sections provide a detailed description of each KACM's component.

Vision-language model. Combining data modalities (i.e., text and image for our use case) is an open challenge in machine learning research and has driven the interest into new (multimodal) tasks, including visual Q&A [55] and topic learning [56]. Within these contexts, novel and more powerful visual-language models have been designed to capture the inner correlations between texts and images [9,11,57].

Table 1
Distributions of datasets. For Memotion7k, we report in parentheses the number of memes that do not convey that affect signal.

Dataset	Class	Train	Validation	Test	Total	Metric
HM	Hateful Non-Hateful	3,050 5,450	340 200	1,250 750	10,040	AUC
MAMI	Hateful Non-Hateful	4,000 4,000	250 250	750 750	10,000	Macro F1-score
MultiOFF	Hateful Non-Hateful	187 258	58 91	58 91	743	Weighted F1-score
Memotion7k	Humour Sarcasm Offensive Motivation	4,272 (1,320) 4,358 (1,236) 3,432 (2,170) 1,973 (3,621)	534 (166) 544 (155) 426 (269) 246 (453)	534 (166) 544 (155) 426 (269) 246 (453)	6,992 6,992 6,992 6,992	Macro F1-score

Under these premises, we leverage the Visual Transformer (ViT) architecture which jointly encodes the image and the text of the meme into a shared representation. Specifically, we adopt the ConcatBERT model [58] which extracts features from images using ResNet-152 and features from text using BERT. To capture inter-modality interactions and learn the joint representation, we incorporate a cross-modal attention layer [9]. Notably, this choice is also driven by the promising results similar models have achieved for disinformation mining tasks, including fake news detection in news articles [59] and social media posts [60].

While acknowledging that there are newer methods for integrating text and image modalities, our work's contribution lies in how we incorporate external knowledge in the classification process rather than solely focusing on exploiting the internal content of memes. Furthermore, the selection of the Visual-Language model is parametric to the KACM architecture, and we will evaluate its impact through experiments.

Sampling strategy. As noted earlier, the knowledge network $\widetilde{\mathcal{G}}_M$ can contain a vast amount of information. For this reason, we define a *bucket* as an elementary unit of knowledge that may (or may not) be informative for the hateful classification task. To implement the concept of the *bucket*, we utilize a random walk approach on $\widetilde{\mathcal{G}}_M$. This process generates a sequence of nodes that constitute the bucket, denoted as $b_i = \{s_1, s_2, \dots, s_l | s_i \in \widetilde{\mathcal{G}}_M\}$, l being the length of the random walk. In particular, we adopt the standard random walk with restart (RWR) algorithm [61], which has demonstrated its effectiveness in various classification tasks [62–64]. To limit the amount of information to be processed, we adopt a *sampling strategy* that selects a predetermined number of k buckets (i.e., random walks) from $\widetilde{\mathcal{G}}_M$ to be used as the memory slots during the classification process.

Pooling. The *pooling component* plays a crucial role in generating a memory unit, which serves as a compact representation of a knowledge *bucket*. In particular, we consider the embeddings of the nodes that belong to the bucket, and perform the mean pooling operation to aggregate them into a single vector. This process generates the memory unit $m_i \in \mathcal{R}^n$ that represents the fundamental knowledge element or bucket, which is injected into the classification process. We conjecture that the mean pooling operation does not loose too much information because the bucket size is limited with respect to the size of $\widetilde{\mathcal{G}}_M$.

Memory summarization. The memory units represent different buckets and thus have different utility on the classification process. Indeed, Fig. 4 shows that the extracted knowledge also includes useless and noisy information which should be ruled out before making the final hateful decision. However, we do not have any information about the correct memory unit(s) which should be considered given the input meme. As a result, we utilize the attention mechanism to dynamically learn the most informative context (i.e., the most relevant units) for the hateful classification. Formally, we consider the group of memory units $\{m_1, m_2, \dots, m_k | m_i \in \mathcal{R}^n\}$ and leverage the soft attention mechanism proposed in [65]:

$$u_{it} = tanh(W_w m_{it} + b_w) \tag{1}$$

$$\alpha_{it} = \frac{\exp(u_{it} u_w)}{\sum_{k=1}^{T_x} \exp(u_{ik} u_w)} \tag{2}$$

$$s_i = \sum_t \alpha_{it} m_{it} \tag{3}$$

That is, the ith memory unit m_{it} is projected to the vector u_{it} through a one-layer MLP. Subsequently, α_{it} is computed representing the normalized similarity between u_{it} and the (jointly learnt) context vector u_w. Next, the ith attention score s_i is computed with the dot product of α_i and m_i and represents the contextual importance of the ith memory unit. Finally, we obtain the memory summary $\mathbf{m} \in \mathcal{R}^n$ as the linear combination of the attention scores and the memory units:

$$\mathbf{m} = \sum_{i=1}^{k} s_i m_i \tag{4}$$

Notably, the MLP weights W_w, b_w and the context vector u_w are dynamically learnt during the training process, allowing the system to automatically recognize which knowledge buckets are informative for the hateful classification. In other words, the ability to dynamically learn the importance of each bucket enables the system to filter out irrelevant or unhelpful information stored in $\widetilde{\mathcal{G}}_M$.

Classification head. This module is responsible for the ultimate hateful classification. In particular, it involves concatenating the embedding of the input meme $\mathbf{q}$ and the memory summary $\mathbf{m}$, followed by passing the concatenated feature representation through a sequence of fully connected layers, culminating in a final softmax layer. The fully connected layers perform a nonlinear mapping of the input features to a higher-dimensional space, enabling the model to learn complex decision boundaries for the classification task. The final softmax layer normalizes the output probabilities across all classes, yielding the predicted probability distribution over the possible classes (e.g., hateful or not under binary settings).

4. Experiments

4.1. Dataset & metrics

We evaluate KERMIT on four well-established benchmark datasets commonly employed in current literature, namely Facebook Hateful Memes (HM), Multimedia Misogyny Dataset (MAMI), MultiOFF and Memotion7k. Table 1 shows some summary statistics of these datasets.

Facebook Hateful Memes (HM) [4] was developed for the homonym challenge and is a pioneering resource for detecting hateful memes. It focuses on identifying hate speech targeting protected categories like race and gender. In particular, we focus on the subset of data used for the second phase of the challenge [33], comprising 8500 training, 540 validation, and 2000 test memes. Annotation was performed according to specific guidelines for consistency. The primary evaluation metric is the Area Under the ROC Curve (AUC).

Multimedia Misogyny (MAMI) [5], from SemEval 2022 Task-5, concentrates on misogynistic content from social media, comprising

10,000 memes from Twitter and Reddit. It utilizes two annotation schemes: binary classification (misogynous vs. non-misogynous) and a 5-class system (general misogyny, shaming, stereotype, objectification, violence). Due to low annotator agreement in the multi-class scheme [5], our focus is on binary classification, with 500 memes per class. Performance is measured using the macro average F1-score.

MultiOFF [26] includes 743 memes from the 2016 U.S. presidential election, categorized as offensive or non-offensive. Following [14], we consider offensive memes as harmful and adopt the weighted F1-score as evaluation metric.

Memotion7k [27], from the SemEval-2020 "Memotion Analysis" task, includes 6992 manually annotated memes for sentiment and affect signals. We focus on the multi-label affect classification, considering combinations of Humorous, Sarcasm, Offensive, and Motivation categories. Performance is measured using the macro average F1-score for each affect signal.

4.2. Experimental setup

4.2.1. Implementation details

All experiments have been performed on Google Colab equipped with one single core hyper threaded Xeon Processor @2.2 GHz, 12 GB of RAM and a Tesla K80 GPU. The code will be made available on Github.[3]

In order to construct the meme's knowledge-enriched information network $\widetilde{\mathcal{G}}_M$, we employ several tools and libraries. For image captioning, we utilize the "Salesforce/blip-image-captioning-base" model from the HuggingFace library.[4] To extract text from the meme's image, we utilize the easyOCR library.[5] Both the embedded text and the caption are processed to remove punctuation and stop words. Next, we employ the Spacy dependency parser[6] to extract the dependency tree of the meme's caption and embedded text. Finally, to retrieve any relevant information from ConceptNet, we leverage the official Python package.[7]

Regarding the KERM module, we adopt the "bert-base-uncased" model from the HuggingFace library[2] to embed the text, and we employ the HIN2Vec implementation developed by [66] for the embedding of the meme's knowledge enriched information network. Additionally, for the KACM module, we used the ConcatBERT model from the MMF library[8] as the base vision-language model. Then, the classification head was implemented with the Pytorch library as a feed-forward neural network with four linear layers, one dropout layer, and the softmax activation function for the final binary classification.

In terms of training, each dataset is trained independently when assessing the impact of knowledge (Section 4.3.2) and evaluating different components of KERMIT's architecture (Section 4.3.3). For the comparative analysis with baseline models (Section 4.3.1), a data augmentation strategy is applied, merging the training sets of the HM and MAMI datasets to fully exploit the available data. This approach aims to maximize the utility of available data as misogynistic content inherently possesses offensive elements. The models are trained for 50 epochs with a batch size of 32, using cross-entropy loss in all training runs.

[3] https://github.com/valeriolagatta/KERMIT_MemeDetection
[4] https://huggingface.co/Salesforce/blip-image-captioning-base
[5] https://pypi.org/project/easyocr/
[6] https://spacy.io/api/dependencyparser
[7] https://pypi.org/project/ConceptNet/
[8] https://github.com/facebookresearch/mmf

4.2.2. Baselines

In our study, we have selected a diverse range of multimodal baseline models, categorized into two distinct groups: (i) models pre-trained on single modalities, such as those combining a pre-trained BERT model [52] with a pre-trained ResNet model [67], and (ii) models pre-trained with multimodal objectives [68]. Specifically, our analysis includes eight baseline models: (1) *VilBERT* [69] processes visual and textual inputs separately, with interaction facilitated through co-attentional transformer layers; (2) *VisualBERT* [70] relies on self-attention within transformers [71] for implicit alignment of text and image elements; (3) *SEER* [72] is pre-trained with diverse web images, aiming to reduce biases; (4) *VisualBERT COCO* [70] is trained on the COCO dataset [73] with a focus on conceptual captions; (5) *ViLT* [74] features a convolution-free architecture for handling text and image modalities; (6) *OSCAR* [75] utilizes object tags in images as anchors for learning alignment with text; (7) *CLIP* [10] employs contrastive learning objectives; (8) *Ernie-Vil* [76] incorporates knowledge from scene graphs [77] for vision-language modeling. Notably, the latter baseline explicitly uses some knowledge to improve the learning process, even if that knowledge is internal to the content rather than external.

For the Memotion7k dataset, which focuses on broader affect signals in memes (i.e., humor, sarcasm, offensiveness, and motivation), KERMIT's performance is compared against baseline models from the SemEval 2020 "Memotion Analysis" task leaderboard [27]. These models integrate features from both image and text modalities [78–80], and some employ multi-task learning [81] and ensemble strategies [82,83].

Finally, we include four state-of-the-art models that are specifically tailored for meme analysis: (1) *MHA-Meme* [37] leverages static word embeddings and trains a LSTM from scratch; (2) *DisMultiHate* [14] focuses on disentangling target information from memes; (3) *MemeFier* [15] employs two fusion strategies to learn inter-modality correlation at different granularity; (4) *PromptHate* [38] reframes the task using model-prompting techniques. While *MHA-Meme* is specifically designed for Memotion7k's multi-label affect classification, the other models are developed for detecting harmful memes.

4.3. Results

4.3.1. Comparison with baselines

Table 2 presents a comparative analysis of KERMIT's performance against baselines and state-of-the-art models for the HM, MAMI, and MultiOFF datasets. The results indicate a general trend wherein models with multimodal pre-training exhibit superior performance compared to those utilizing unimodal pre-training. This aligns with existing research [84], underscoring the significance of multimodal interactions in meme comprehension for enhanced downstream task performance. Notably, Ernie-ViL demonstrates marginally better results on HM and MAMI datasets, suggesting the beneficial impact of integrating meme's internal knowledge into the classification process. However, KERMIT surpasses Ernie-ViL by 5.5%, 4.9%, and 18.4% on the HM, MAMI, and MultiOFF datasets, respectively, emphasizing the advantages of explicitly integrating external knowledge.

Regarding state-of-the-art competitors, MHA-Meme shows lower performance across datasets, possibly due to its approach of training from scratch a sequence model instead of leveraging pre-trained models. The performance of DisMultiHate, MemeFier, and PromptHate varies across datasets, with no single model consistently excelling. We believe that their success, relative to baselines, can be attributed to the integration of external knowledge, including image tags (e.g., web entities) and demographic information (e.g., race, gender, age). On the other hand, their custom training mechanism and modality fusion strategy determine their relative performance gap but this contribution seems marginal and largely depends on the actual data at hand. Notably, PromptHate's underperformance on the MultiOFF dataset may be due to limitations in handling extensive text input within a single prompt. Indeed, MultiOFF memes are characterized by longer texts (45

Table 2
Comparison with baselines and state-of-the-art performers (bold indicates the best results, underline the first runner up).

Category	Model	Dataset		
		HM	MAMI	MultiOFF
Multimodal (Unimodal pre-training)	VilBERT [69]	0.734	0.725	0.557
	VisualBERT [70]	0.732	0.723	0.562
	SEER [72]	0.708	0.718	0.543
Multimodal (Multimodal pre-training)	ViLT [74]	0.725	0.744	0.623
	VisualBERT COCO [70]	0.752	0.742	0.577
	OSCAR [75]	0.793	0.684	0.606
	CLIP [10]	0.803	0.765	0.617
	Ernie-Vil [76]	0.806	0.793	0.531
SOTA Competitors	MHA-Meme [37]	0.658	0.688	0.591
	DisMultiHate [14]	0.799	0.801	<u>0.643</u>
	MemeFier [15]	0.801	<u>0.832</u>	0.621
	PromptHate [38]	<u>0.814</u>	0.799	0.420
	KERMIT (Ours)	**0.853**	**0.834**	**0.651**
	Δ_{KERMIT} (%)	4.3%	0.3%	1.3%

Table 3
Comparison with competitors at Semeval 2020 "Memotion Analysis" task and state-of-the-art approaches for the Memotion7k dataset (bold indicates the best results, underline the first runner up).

Category	Model	Humour	Sarcasm	Offensive	Motivation	Average
Memotion Analysis	NUAA [78]	0.434	0.447	0.400	0.488	0.442
	IITK [85]	0.473	0.508	0.499	0.473	0.488
	Walińska [83]	0.502	0.499	0.479	0.498	0.494
	CSECU [79]	0.493	0.487	0.505	0.490	0.494
	Memebusters [86]	0.529	0.485	0.529	0.491	0.508
	PRHLT [80]	0.510	0.513	0.506	0.509	0.509
	Guoym [82]	0.515	0.511	0.512	0.520	0.515
	UPB [81]	0.516	<u>0.516</u>	0.522	0.519	0.518
SOTA Competitor	MHA-Meme [37]	0.527	**0.520**	0.517	0.531	<u>0.523</u>
	Memefier [15]	**0.549**	0.451	<u>0.529</u>	**0.543**	0.518
	KERMIT (Ours)	<u>0.538</u>	0.495	**0.535**	<u>0.531</u>	**0.525**
	Δ_{KERMIT}	−2.0%	−4.8%	1.1%	−2.2%	0.4%

words per meme on average) with respect to the texts in the memes of other datasets (16 words per meme on average).

Overall, KERMIT consistently achieves promising results across all datasets, with significant improvement on the HM dataset and marginal gains on MAMI and MultiOFF. This underscores the importance of knowledge modeling strategy in terms of the meme's *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$. Please refer to the next section for additional details about the effects on knowledge modeling strategy.

On Memotion7k, as shown in Table 3, the performance gap between "Memotion Analysis" task competitors and state-of-the-art models is narrower. This could be due to the dataset's multi-label nature and significant class imbalance. Despite these challenges, KERMIT competes closely with top models and notably excels in the "offensive" affect signal category, which is the closest to the harmful category that our system targets. Also, KERMIT slightly improves the average performance of MHA-Meme by 0.4%.

In summary, these experimental results validate KERMIT as an effective approach for classifying harmful memes, irrespective of the type of harm or affect signal. The integration of external knowledge into the model architecture not only positions KERMIT favorably against various benchmarks but also suggests that incorporating such knowledge could be a vital direction for future multimodal disinformation detection research.

4.3.2. What is the contribution of the external knowledge?

The primary hypothesis of this study is that comprehending harmful memes necessitates background knowledge that conveys the context of the meme, such as cultural allusions and societal concerns. Consequently, we first investigate the impact of incorporating knowledge into the classification process. Specifically, we explore four different settings: (i) *no knowledge*, i.e., a scenario without any knowledge, where only the vision-language model and the classification head in the KACM module are evaluated; (ii) *raw knowledge*, i.e., a scenario with raw knowledge, where the caption and embedded text are combined with the raw terms extracted from ConceptNet and then, all together, are fed to vision-language model and the classification head; (iii) *ConceptNet*, i.e., a scenario where the meme's *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$ is built using external knowledge obtained from ConceptNet [24] as described in Section 3.2.1; and (iv) *Wikidata*, i.e., this setting is similar to the previous one, except that $\widetilde{\mathcal{G}}_M$ is built using WikiData [51] instead of ConceptNet.

The histograms in Fig. 6a indicates that classification performance improves under both the *ConceptNet* and *WikiData* settings, compared to the *no knowledge* scenario. Particularly, the performance gains in the Memotion7k dataset are notable, with increases (over the *no knowledge* scenario) of 17.5%, 26.9%, 27.3%, and 5.41% for humor, motivation, offensive, and sarcasm affect signals, respectively, when external knowledge from *ConceptNet* is utilized. This finding validates our hypothesis that the inclusion of external knowledge positively influences the classification process, independent of the knowledge source's nature or the specific entities it involves.

Nonetheless, we also emphasize that, except for the "offensive" category in Memotion7k, the *ConceptNet* settings exhibits better performance relative to the *WikiData* settings, indicating that the commonsense knowledge incorporated in ConceptNet might be more appropriate than the purely factual knowledge contained in WikiData. Furthermore, the *raw knowledge* setting does not consistently lead to performance gains, as evidenced for the HM dataset and the "sarcasm" affect signal in Memotion7k. This result provides empirical evidence supporting the efficacy of the knowledge modeling strategy introduced in Section 3.2.1. It implies that merely incorporating unstructured knowledge is insufficient for enhancing classification performance. Instead, structuring knowledge to enable the model to effectively extract

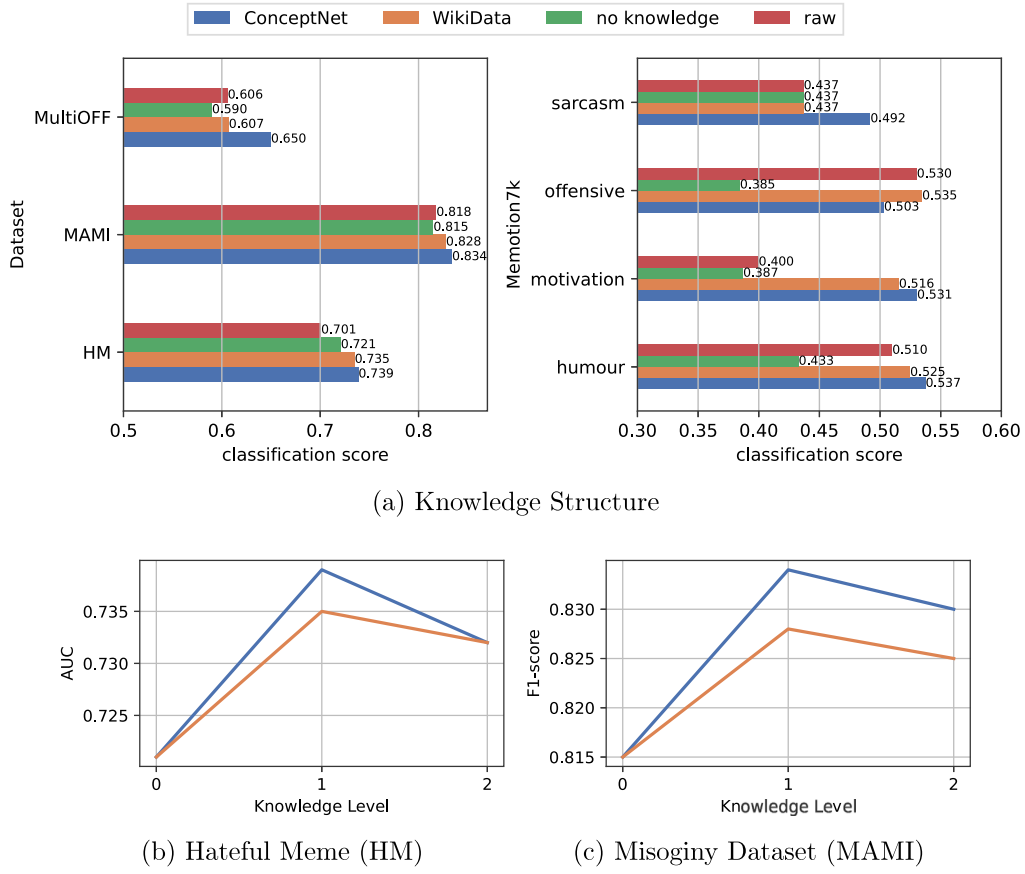

Fig. 6. The contribution of external knowledge: (a) comparison between structured knowledge from ConceptNet, structured knowledge from WikiData, unstructured knowledge and no knowledge; (b), (c) performance by varying the recursion depth on Hateful Meme and MAMI datsets, respectively.

pertinent background information is critical for identifying harmful signals in memes.

To further investigate the advantages of incorporating external knowledge, we examine the "quantity" of knowledge required to improve classification performance. Specifically, we examined the optimal depth of knowledge integration with respect to classification performance on HM and MAMI datasets. Figs. 6b and 6c show the impact of varying recursion depth when querying ConceptNet (or WikiData). Although performance differences between recursion depths of one ($l = 1$) and two ($l = 2$) are modest, we posit that a depth of one is optimal. This conclusion is based on two considerations: first, the relevance of knowledge to the meme tends to diminish with increased recursion depth, elevating the risk of integrating irrelevant or noisy information into $\widetilde{\mathcal{G}}_M$; second, a higher recursion depth exponentially increases the number of API queries to the external knowledge base, leading to impractical construction times for $\widetilde{\mathcal{G}}_M$. These insights emphasize the delicate balance required between depth of knowledge integration and practical computational limitations.

It is worth to highlight a variance in performance on the HM dataset in Fig. 6, with respect to the results reported in Table 2. In contrast, performance on MAMI remains consistent. This observed discrepancy is attributed to the methodological differences in dataset handling between experiments. Specifically, in this experiment, datasets are analyzed independently, as opposed to the merging of HM and MAMI training sets utilized in the previous setup. The underlying reason for this differential impact lies in the hierarchical nature of the tasks associated with the HM and MAMI datasets. Indeed, while misogynistic content typically encompasses offensive elements, not all offensive content necessarily contains misogynistic elements. Therefore, the inclusion of misogynistic memes from the MAMI dataset contributed to enhancing the detection of harmful content in the HM dataset, whereas general harmful content did not similarly aid in detecting misogyny.

Overall, these findings highlight the advantages of incorporating external knowledge within the classification process of harmful memes. Furthermore, our results indicate that the *knowledge-enriched information network* constructed by KERMIT as well as the dynamic learning mechanism designed in its KACM module are capable of effectively capturing critical information about the input meme.

4.3.3. Ablation study

Once having established the advantages of adding external knowledge, we now shift our evaluation targets to the other components of KERMIT, i.e., the vision-language model in the KACM module, the node embedding algorithm in the KERM module, the knowledge injection strategy in the KACM module.

Vision-language model. We consider the vision-language model of the KACM module. As noted earlier, this module aims to represent the internal information conveyed by the meme and capture the joint relationship between the image and text. To assess the performance of different vision-language models, we conduct experiments utilizing two distinct models: ConcatBERT [58] and MMBT [11]. Results in Table 4 indicate that both models demonstrate comparable effectiveness, with ConcatBERT showing marginally superior performance on the HM dataset. Crucially, the integration of the meme's *knowledge-enriched information network* $\widetilde{\mathcal{G}}_M$ into the classification process yielded notable performance improvements for both ConcatBERT and MMBT. This finding suggests that the *knowledge-enriched information network* captures information that is not captured by the contextual embedding of vision-language models and highlights their complementary utility to improve detection performance.

Table 4
Ablation study: performance by varying the vision-language model in the KACM module, the node embedding algorithm in the KERM module, the knowledge injection strategy in the KACM module.

Module	Parameter	Dataset HM	MAMI
Visual-Language Model	ConcatBERT	0.721	0.815
	+ $\widetilde{\mathcal{G}}_M$	**0.739**	**0.834**
	MMBT	0.718	0.815
	+ $\widetilde{\mathcal{G}}_M$	**0.729**	**0.817**
KERM	FeatherNode	0.708	0.806
	Hin2Vec	**0.739**	**0.834**
KACM	Average	0.725	0.825
	Attention	**0.739**	**0.834**

Graph embedding. We investigate the impact of the node embedding algorithm used in the KERM module. Specifically, we compare the performance of HIN2Vec [53] and FeatherNode [87] algorithms, as shown in Table 4. Our results reveal that HIN2Vec outperforms FeatherNode, possibly due to its higher representational capacity. Indeed, while FeatherNode treats the $\widetilde{\mathcal{G}}_M$ as a homogeneous graph and neglects the semantics of relationships between nodes, HIN2Vec is tailored to heterogeneous networks and leverages the relationship types among nodes. Thus, the learnt embeddings take into account the node types (i.e., whether a node represents a meme's entity or a concept retrieved from ConceptNet) as well as the relationship types between nodes.

Knowledge injection. We investigate the impact of two design choices on the performance of the KERMIT model, namely the size of knowledge buckets b_i and the memory summarization strategy. The former parameter is examined by varying the number and length of random walks within the model. Our experimental findings, as illustrated in Fig. 7, suggest that these parameters exert a relatively modest impact on model's performance. This outcome probably depends on two contrasting trends related to the size variance of $\widetilde{\mathcal{G}}_M$ within the same dataset. Specifically, while memes linked to larger graphs may benefit from more extensive knowledge buckets, those associated with smaller graphs could experience a decrease in performance due to the introduction of noisy knowledge. Nevertheless, we find that optimal results for both datasets is consistently achieved with a configuration of 15 random walks. Furthermore, we determine that a random walk length of 11 nodes is preferable since increasing this parameter also comes with computational cost. Overall, this configuration not only enhances performance but also improves computational efficiency, a key consideration in the practical deployment of the model.

Finally, to evaluate the effectiveness of the memory summarization technique used in the KACM module, we conducted a comparison between the attention mechanism proposed in Section 3.2.3 and a simpler method that averages the embeddings of each memory unit. The results are presented in Table 4 and reveal that the attention-based approach achieves superior classification performance. This result confirms our hypothesis that memory units do not have the same impact on hateful classification and indicates that the model is capable of (dynamically) learning the most informative buckets without any supervision.

5. Conclusions and future works

In this paper, we presented KERMIT (Knowledge-EmpoweRed Model In harmful meme deTection), a novel framework designed for knowledge-informed harmful meme detection. The proposed approach involves the creation of a knowledge-enriched information network for memes by integrating internal entities of the meme with relevant external knowledge obtained from ConceptNet. KERMIT is further empowered with a dynamic learning mechanism, utilizing memory-augmented neural networks and attention mechanism to discern the most informative knowledge for accurate classification of harmful memes. The effectiveness of the proposed approach was demonstrated through experiments conducted on four benchmark datasets, i.e., Facebook Hateful Memes, mysoginy dataset (MAMI), MultiOFF and Memotion7k. Specifically, the results showed that KERMIT is capable of retrieving contextual knowledge and utilizing it effectively in the classification process to enhance predictive performance. Moreover, when compared to several state-of-the-art competitors in the field of harmful meme detection, KERMIT achieved comparable or superior classification performance across all datasets.

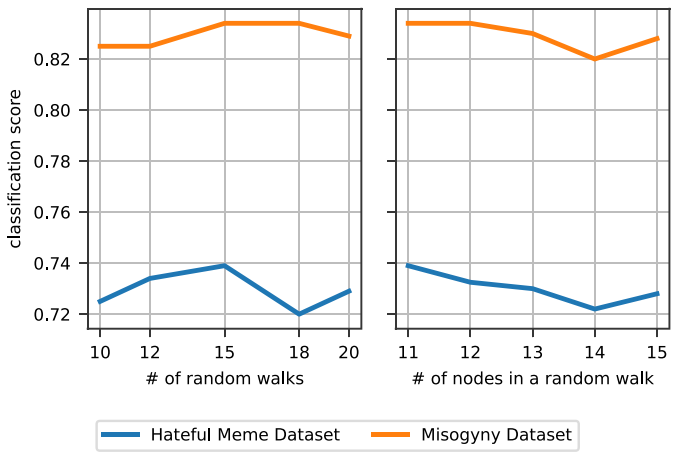

Fig. 7. Ablation study: performance by varying the size of the knowledge buckets.

In summary, our study underscores the efficacy of integrating external knowledge into the classification process and paves the way for further exploration in the domain of harmful meme detection. The findings accentuate the crucial role of artificial intelligence and knowledge discovery in advancing content moderation.

There is a number of avenues for future works that we would like to explore. First, we plan to extend our methodology to dynamically tailor the knowledge integration process, such as adjusting the number or size of knowledge buckets, to the specific characteristics of each meme. Second, we plan to explore the use of multimodal knowledge bases, such as VisualGenome [88], to extract external knowledge. Currently, KERMIT focuses solely on textual knowledge, but we believe that incorporating visual knowledge will enhance the detection of harmful memes. Third, we would like to utilize the meme's *knowledge-enriched information network* to provide interpretability power to our framework, i.e., we expect that the extracted knowledge can help us understand why (or why not) a particular meme is harmful. In this way, we can provide more transparency to the classification process, which can aid in content moderation. Finally, we aim to explore the ethical implications of KERMIT's use, particularly in relation to potential biases or prejudices that may be exacerbated by the addition of external knowledge.

CRediT authorship contribution statement

Biagio Grasso: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Valerio La Gatta:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Vincenzo Moscato:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Giancarlo Sperlì:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are open-source. The code will be made available on GitHub.

Acknowledgments

This work has been funded by PNRR MUR, Italy project PE0000013-FAIR and by the Spoke 9 "Digital Society & Smart Cities" of ICSC - Centro Nazionale di Ricerca in High Performance-Computing, Big Data and Quantum Computing, PNRR-HPC, Italy (CUP: E63C22000980007).

References

[1] L. Glitsos, J. Hall, The pepe the frog meme: An examination of social, political, and cultural implications through the tradition of the Darwinian Absurd, J. Cult. Res. 23 (4) (2019) 381–395.

[2] C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, B. Gambäck, SemEval-2020 task 8: Memotion analysis - the visuo-lingual metaphor!, 2020, CoRR, arXiv:2008.03781, [Online]. Available: https://arxiv.org/abs/2008.03781.

[3] D. Dimitrov, B.B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, G.D.S. Martino, SemEval-2021 task 6: Detection of persuasion techniques in texts and images, 2021, CoRR, arXiv:2105.09284, [Online]. Available: https://arxiv.org/abs/2105.09284.

[4] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[5] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation, (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549, [Online]. Available: https://aclanthology.org/2022.semeval-1.74.

[6] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, G. Suarez-Tangil, On the origins of memes by means of fringe web communities, in: Proceedings of the Internet Measurement Conference 2018, IMC '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 188–202.

[7] R. Zhu, Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution, 2020, CoRR, arXiv:2012.08290, [Online]. Available: https://arxiv.org/abs/2012.08290.

[8] S. Sharma, F. Alam, M.S. Akhtar, D. Dimitrov, G. Da San Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, T. Chakraborty, Detecting and understanding harmful memes: A survey, in: L.D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5597–5606, Survey Track.

[9] L.H. Li, M. Yatskar, D. Yin, C. Hsieh, K. Chang, VisualBERT: A simple and performant baseline for vision and language, 2019, CoRR, arXiv:1908.03557, [Online]. Available: http://arxiv.org/abs/1908.03557.

[10] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021, CoRR, arXiv:2103.00020, [Online]. Available: https://arxiv.org/abs/2103.00020.

[11] D. Kiela, S. Bhooshan, H. Firooz, D. Testuggine, Supervised multimodal bitransformers for classifying images and text, 2019, CoRR, arXiv:1909.02950, [Online]. Available: http://arxiv.org/abs/1909.02950.

[12] B. Kostadinovska-Stojchevska, E. Shalevska, Internet memes and their socio-linguistic features, Eur. J. Lit., Lang. Linguist. Stud. 2 (4) (2018).

[13] C.-C. Lin, Y.-C. Huang, J.Y.-j. Hsu, Crowdsourced explanations for humorous internet memes based on linguistic theories, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 2, (1) 2014, pp. 143–150, [Online]. Available: https://ojs.aaai.org/index.php/HCOMP/article/view/13169.

[14] R.K.-W. Lee, R. Cao, Z. Fan, J. Jiang, W.-H. Chong, Disentangling hate in online memes, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5138–5147.

[15] C. Koutlis, M. Schinas, S. Papadopoulos, MemeFier: Dual-stage modality fusion for image meme classification, in: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 586–591.

[16] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, B. Faieta, Multimodal contrastive training for visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 6995–7004.

[17] H. Tan, M. 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, Hong Kong, China, 2019, pp. 5100–5111, [Online]. Available: https://aclanthology.org/D19-1514.

[18] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, D. Kiela, FLAVA: A foundational language and vision alignment model, 2021, CoRR, arXiv:2112.04482, [Online]. Available: https://arxiv.org/abs/2112.04482.

[19] A. Silva, P. Tambwekar, M. Gombolay, Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2383–2389, [Online]. Available: https://aclanthology.org/2021.naacl-main.189.

[20] A.K. Thakur, F. Ilievski, H.-A. Sandlin, Z. Sourati, L. Luceri, R. Tommasini, A. Mermoud, Multimodal and explainable internet meme classification, 2023, arXiv:2212.05612.

[21] M.S. Hee, R.K.-W. Lee, W.-H. Chong, On explaining multimodal hateful meme detection models, in: Proceedings of the ACM Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 3651–3655.

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

[23] E. Ferrara, Should ChatGPT be biased? Challenges and risks of bias in large language models, First Monday 28 (11) (2023) [Online]. Available: https://firstmonday.org/ojs/index.php/fm/article/view/13346.

[24] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, 2016, CoRR, arXiv:1612.03975, [Online]. Available: http://arxiv.org/abs/1612.03975.

[25] L. Shang, C. Youn, Y. Zha, Y. Zhang, D. Wang, KnowMeme: A knowledge-enriched graph neural network solution to offensive meme detection, in: 2021 IEEE 17th International Conference on EScience, (EScience), 2021, pp. 186–195.

[26] S. Suryawanshi, B.R. Chakravarthi, M. Arcan, P. Buitelaar, Multimodal meme dataset (multiOFF) for identifying offensive content in image and text, in: R. Kumar, A.K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, D. Kadar (Eds.), Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 32–41, [Online]. Available: https://aclanthology.org/2020.trac-1.6.

[27] C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, B. Gambäck, SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor!, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 759–773, [Online]. Available: https://aclanthology.org/2020.semeval-1.99.

[28] S. Pramanick, S. Sharma, D. Dimitrov, M.S. Akhtar, P. Nakov, T. Chakraborty, MOMENTA: A multimodal framework for detecting harmful memes and their targets, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4439–4455, [Online]. Available: https://aclanthology.org/2021.findings-emnlp.379.

[29] S. Sharma, M.S. Akhtar, P. Nakov, T. Chakraborty, DISARM: Detecting the victims targeted by harmful memes, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1572–1588, [Online]. Available: https://aclanthology.org/2022.findings-naacl.118.

[30] S. Sharma, A. Kulkarni, T. Suresh, H. Mathur, P. Nakov, M.S. Akhtar, T. Chakraborty, Characterizing the entities in harmful memes: Who is the hero, the villain, the victim?, 2023, arXiv:2301.11219.

[31] L. Piras, G. Giacinto, Information fusion in content based image retrieval: A comprehensive overview, Inf. Fusion 37 (2017) 50–60, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517300076.

[32] D. Zhang, R. Cao, S. Wu, Information fusion in visual question answering: A survey, Inf. Fusion 52 (2019) 268–280, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253518308893.

[33] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, C.A. Fitzpatrick, P. Bull, G. Lipstein, T. Nelli, R. Zhu, N. Muennighoff, R. Velioglu, J. Rose, P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, H. Yannakoudakis, V. Sandulescu, U. Ozertem, P. Pantel, L. Specia, D. Parikh, The hateful memes challenge: Competition report, in: H.J. Escalante, K. Hofmann (Eds.), Proceedings of the NeurIPS 2020 Competition and Demonstration Track, in: Proceedings of Machine Learning Research, vol. 133, PMLR, 2021, pp. 344–360, [Online]. Available: https://proceedings.mlr.press/v133/kiela21a.html.

[34] J. Zhang, Y. Wang, SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider,

S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation, (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 585–596, [Online]. Available: https://aclanthology.org/2022.semeval-1.81.

[35] R. Velioglu, J. Rose, Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge, 2020, CoRR, arXiv:2012.12975, [Online]. Available: https://arxiv.org/abs/2012.12975.

[36] G.K. Kumar, K. Nandakumar, Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features, in: L. Biester, D. Demszky, Z. Jin, M. Sachan, J. Tetreault, S. Wilson, L. Xiao, J. Zhao (Eds.), Proceedings of the Second Workshop on NLP for Positive Impact, (NLP4PI), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 171–183, [Online]. Available: https://aclanthology.org/2022.nlp4pi-1.20.

[37] S. Pramanick, M.S. Akhtar, T. Chakraborty, Exercise? I thought you said 'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 15, (1) 2021, pp. 513–524, [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/18080.

[38] R. Cao, R.K.-W. Lee, W.-H. Chong, J. Jiang, Prompting for multimodal hateful meme classification, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 321–332, [Online]. Available: https://aclanthology.org/2022.emnlp-main.22.

[39] R. Cao, M.S. Hee, A. Kuek, W.-H. Chong, R.K.-W. Lee, J. Jiang, Pro-cap: Leveraging a frozen vision-language model for hateful meme detection, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5244–5252.

[40] J. Weston, S. Chopra, A. Bordes, Memory networks, 2015, Publisher Copyright: © 2015 International Conference on Learning Representations, ICLR. All rights reserved.; 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

[41] C. Ma, C. Shen, A. Dick, Q. Wu, P. Wang, A.v.d. Hengel, I. Reid, Visual question answering with memory-augmented networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6975–6984.

[42] Z. Fei, Memory-augmented image captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, (2) 2021, pp. 1317–1324, [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16220.

[43] E. Parisotto, R. Salakhutdinov, Neural map: Structured memory for deep reinforcement learning, in: International Conference on Learning Representations, 2018, [Online]. Available: https://openreview.net/forum?id=Bk9zbyZCZ.

[44] F. Ruggeri, M. Lippi, P. Torroni, Membert: Injecting unstructured knowledge into BERT, 2021, CoRR, arXiv:2110.00125, [Online]. Available: https://arxiv.org/abs/2110.00125.

[45] E. Grefenstette, K.M. Hermann, M. Suleyman, P. Blunsom, Learning to transduce with unbounded memory, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Vol. 2, NIPS '15, MIT Press, Cambridge, MA, USA, 2015, pp. 1828–1836.

[46] S. Moon, P. Shah, A. Kumar, R. Subba, Memory graph networks for explainable memory-grounded question answering, in: Proceedings of the 23rd Conference on Computational Natural Language Learning, (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 728–736, [Online]. Available: https://aclanthology.org/K19-1068.

[47] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, pp. 12888–12900, [Online]. Available: https://proceedings.mlr.press/v162/li22n.html.

[48] L. Cui, F. Wei, M. Zhou, Neural open information extraction, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 407–413, [Online]. Available: https://aclanthology.org/P18-2065.

[49] J. Fan, A. Kalyanpur, D.C. Gondek, D.A. Ferrucci, Automatic knowledge extraction from documents, IBM J. Res. Dev. 56 (3.4) (2012) 5:1–5:10.

[50] A. Rossanez, J.C. Dos Reis, R.d.S. Torres, H. de Ribaupierre, Kgen: A knowledge graph generator from biomedical scientific literature, BMC Med. Inform. Decis. Mak. 20 (4) (2020) 1–24.

[51] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85.

[52] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, 2018, CoRR, arXiv:1810.04805, [Online]. Available: http://arxiv.org/abs/1810.04805.

[53] T.-y. Fu, W.-C. Lee, Z. Lei, HIN2vec: Explore meta-paths in heterogeneous information networks for representation learning, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, 2017, pp. 1797–1806.

[54] S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, End-to-end memory networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Vol. 2, NIPS '15, MIT Press, Cambridge, MA, USA, 2015, pp. 2440–2448.

[55] Y. Srivastava, V. Murali, S.R. Dubey, S. Mukherjee, Visual question answering using deep learning: A survey and performance analysis, 2019, CoRR, arXiv:1909.01860, [Online]. Available: http://arxiv.org/abs/1909.01860.

[56] Y. Li, M. Yang, Z. Zhang, A survey of multi-view representation learning, IEEE Trans. Knowl. Data Eng. 31 (10) (2019) 1863–1883.

[57] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019, CoRR, arXiv:1908.02265, [Online]. Available: http://arxiv.org/abs/1908.02265.

[58] A. Singh, V. Goswami, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, D. Parikh, MMF: A multimodal framework for vision and language research, 2020, https://github.com/facebookresearch/mmf.

[59] Y. Wang, Y. Huang, Y. Yang, Multi-modal transformer for fake news detection, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4090–4100.

[60] Q. Jing, D. Yao, X. Fan, B. Wang, H. Tan, X. Bu, J. Bi, TRANSFAKE: Multi-task transformer for multimodal enhanced fake news detection, in: 2021 International Joint Conference on Neural Networks, IJCNN, 2021, pp. 1–8.

[61] H. Tong, C. Faloutsos, J. Pan, Fast random walk with restart and its applications, in: Proceedings of the 22nd International Conference on Data Engineering, ICDE'06, IEEE, 2006, pp. 613–624.

[62] H. Guo, Y. Liu, J. Wang, Y. Wu, Knowledge graph-based multi-label classification with label embeddings and label dependency, Knowl.-Based Syst. 198 (2020) 105965.

[63] Y. Zhang, Q. Yang, D. Zhou, Multi-label classification via knowledge graph embeddings and soft-constrained label propagation, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18, AAAI Press, 2018, pp. 3350–3356.

[64] X. Zhang, S. Xie, H. Liu, M. Sun, Deep learning based recommender system: A survey and new perspectives, ACM Comput. Surv. 52 (1) (2018) 1–38.

[65] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: HLT-NAACL, 2016.

[66] C. Yang, Y. Xiao, Y. Zhang, Y. Sun, J. Han, Heterogeneous network representation learning: A unified framework with survey and benchmark, TKDE (2020).

[67] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015, CoRR, arXiv:1512.03385, [Online]. Available: http://arxiv.org/abs/1512.03385.

[68] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, S. Soatto, Masked vision and language modeling for multi-modal representation learning, in: The Eleventh International Conference on Learning Representations, 2023, [Online]. Available: https://openreview.net/forum?id=ZhuXksSJYWn.

[69] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2019.

[70] L.H. Li, M. Yatskar, D. Yin, C. Hsieh, K. Chang, Visualbert: A simple and performant baseline for vision and language, 2019, arXiv, arXiv preprint arXiv:1908.03557.

[71] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[72] P. Goyal, Q. Duval, I. Seessel, M. Caron, I. Misra, L. Sagun, A. Joulin, P. Bojanowski, Vision models are more robust and fair when pretrained on uncurated images without supervision, 2022, arXiv:2202.08360.

[73] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[74] W. Kim, B. Son, I. Kim, ViLT: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, 2021.

[75] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-semantics aligned pre-training for vision-language tasks, in: ECCV 2020, 2020.

[76] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, H. Wang, ERNIE-vil: Knowledge enhanced vision-language representations through scene graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 3208–3216.

[77] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Image retrieval using scene graphs, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 3668–3678.

[78] X. Guo, J. Ma, A. Zubiaga, NUAA-QMUL at SemEval-2020 task 8: Utilizing BERT and DenseNet for internet meme emotion analysis, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 901–907, [Online]. Available: https://aclanthology.org/2020.semeval-1.114.

[79] A.N. Chy, U.A. Siddiqua, M. Aono, CSECU_KDE_MA at SemEval-2020 task 8: A neural attention model for memotion analysis, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1106–1111, [Online]. Available: https://aclanthology.org/2020.semeval-1.146.

[80] G.L. De la Peña Sarracén, P. Rosso, A. Giachanou, PRHLT-UPV at SemEval-2020 task 8: Study of multimodal techniques for memes analysis, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 908–915, [Online]. Available: https://aclanthology.org/2020.semeval-1.115.

[81] G.-A. Vlad, G.-E. Zaharia, D.-C. Cercel, C. Chiru, S. Trausan-Matu, UPB at SemEval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1208–1214, [Online]. Available: https://aclanthology.org/2020.semeval-1.160.

[82] Y. Guo, J. Huang, Y. Dong, M. Xu, Guoym at SemEval-2020 task 8: Ensemble-based classification of visuo-lingual metaphor in memes, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1120–1125, [Online]. Available: https://aclanthology.org/2020.semeval-1.148.

[83] U. Walińska, J. Potoniec, Urszula walińska at SemEval-2020 task 8: Fusion of text and image features using LSTM and VGG16 for memotion analysis, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1215–1220, [Online]. Available: https://aclanthology.org/2020.semeval-1.161.

[84] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, W. Gao, Large-scale multi-modal pre-trained models: A comprehensive survey, 2023, arXiv:2302.10035.

[85] V. Keswani, S. Singh, S. Agarwal, A. Modi, IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of internet memes, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1135–1140, [Online]. Available: https://aclanthology.org/2020.semeval-1.150.

[86] M. Sharma, I. Kandasamy, W. Vasantha, Memebusters at SemEval-2020 task 8: Feature fusion model for sentiment analysis on memes using transfer learning, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1163–1171, [Online]. Available: https://aclanthology.org/2020.semeval-1.154.

[87] B. Rozemberczki, R. Sarkar, Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1325–1334.

[88] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (2017) 32–73.