Computational Meme Understanding: A Survey

Khoi P. N. Nguyen and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
khoi.nguyen6@utdallas.edu, vince@hlt.utdallas.edu

Abstract

Computational Meme Understanding, which concerns the automated comprehension of memes, has garnered interest over the last four years and is facing both substantial opportunities and challenges. We survey this emerging area of research by first introducing a comprehensive taxonomy for memes along three dimensions – forms, functions, and topics. Next, we present three key tasks in Computational Meme Understanding, namely, classification, interpretation, and explanation, and conduct a comprehensive review of existing datasets and models, discussing their limitations. Finally, we highlight the key challenges and recommend avenues for future work.[1]

1 Introduction

In the current age, *memes* — user-created combinations of pictures and images overlaid with text — are widespread due to their nature of being amusing and relatively quick to consume compared to text. Memes have become a novel and prevalent means of online communication (Joshi et al., 2024).

Memes can be malicious, such as those that are hateful, harmful, or politically manipulative. For example, during the last two US presidential elections, a large amount of coordinated media content, especially memes, was used to affect public opinion and influence the election results.[2] However, given the vastness of the internet, it is impossible to have human workers inspect every single meme on the social network platforms (Kiela et al., 2020). Having an automated system to detect malignity in memes hence can be of great help.

At the same time, as memes are often used to express thoughts and personal opinions, we could use meme understanding technologies to help improve human communication. For example, teachers may understand their students' thoughts better via the memes the students composed, thus possibly helping to narrow the generation gap between them. A student who just started their study in a foreign country could adapt to the new culture better if they can understand the jokes made by their friends via the memes they composed.

Such societal motivations have spurred an interest in Computational Meme Understanding (henceforth, CMU), which is an umbrella term we *introduce* to refer to a collection of tasks involving the automated comprehension of memes. CMU presents several key challenges to researchers. First, CMU systems need to seamlessly recognize and combine textual and visual elements from the meme itself. Second, generating a full textual description of the message conveyed in a meme, a key task in CMU, requires both breadth and depth of knowledge about recent news, internet subcultures, meme cultures, and the world. Finally, CMU systems often need to read between the lines, such as decoding figurative language, to successfully understand a meme.

Our goal in this paper is to present a timely and comprehensive review of work on CMU. To our knowledge, there have been no comprehensive surveys in this area of research. Related surveys are mostly *tangentially* related to memes, including those on computational propaganda (Martino et al., 2020; Ng and Li, 2023), multimodal disinformation and fact-checking (Alam et al., 2022; Akhtar et al., 2023), hate speech (Schmidt and Wiegand, 2017), and humor generation (Amin and Burghardt, 2020). The closest review for memes is by Sharma et al. (2022a), which was written two years ago and concerns only *harmful* memes and the associated *classification* tasks. Our review has a broader scope, covering many more types of memes and more technical tasks.

[1] For illustration purposes, this paper includes some graphics that might be offensive to certain readers.

[2] https://www.newyorker.com/news/annals-of-communications/the-meme-ification-of-american-politics

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21251–21267

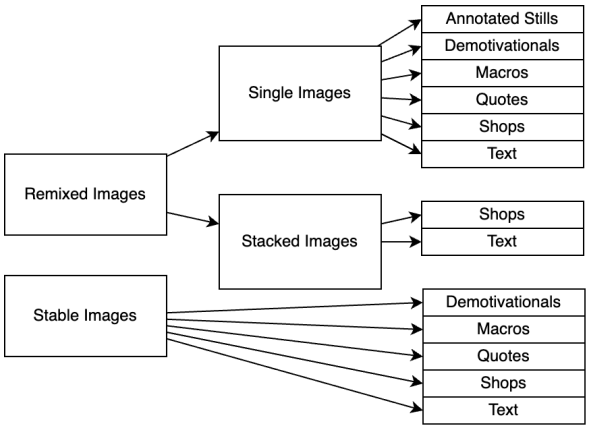


Figure 1: Taxonomy of forms for memes, adapted from Milner (2012).

2 A Taxonomy for Memes

Recognizing the types of memes in a dataset, as well as the distribution of such types in the wild, is important in directing research efforts to align with real-world needs. For example, if a type of memes is popular on the Web, but is not collected, trained, and evaluated on, models may suffer from out-of-distribution problems in production settings.

Before discussing the distribution of meme types, we need a unified language in which a meme can be classified. Currently, meme taxonomies in NLP are narrowly defined within a subset of memes such as the harmful ones (Sharma et al., 2022a; Banko et al., 2020; Arora et al., 2023; Pramanick et al., 2021a). Therefore, we borrow from the social sciences a taxonomy of memes along three dimensions: forms, functions, and topics.

2.1 Forms

Memes have many forms, each of which has distinct ways of creating meaning. Milner (2012), a communication researcher, has developed a widely-cited taxonomy of forms for memes, which is illustrated in Figure 1. In this system, the set of all possible memes can be broadly divided into two groups: *Remixed Images* and *Stable Images*.

Remixed Images are memes created via image manipulation. A remixed image contains a single or multiple sub-images, each of which can be manipulated in various ways. Milner (2012) observed that the most popular type of manipulation was *Macros*, which comprises a base template, a line of text at the top (premise) and another at the bottom (punchline). Another type of manipulation is to add parts of other images to the base image or graphically edit the base image itself — Milner referred to those as *Shops*, which stands for "Photoshop". Other forms of manipulation are *Annotated Stills*, *Demotivationals*, *Quotes*, and *Text*. Finally, multiple remixed images can be stacked together to make a more complex meme.

On the other hand, *Stable Images* are images used as memes without editing. For example, *Screenshots* – e.g., of conversations on social media – can be used as memes. *Photos*, including photos of memes in real life (*Memes IRL*), as well as *Drawings* and *Graphs* are the other categories of *Stable Image* memes.

2.2 Functions

As a means of media and communication, memes are highly functional. Contemporary literature widely considers the default function of memes to be making a joke about something or someone (Milner, 2012; Grundlingh, 2018). On top of that, a meme often does something else, such as persuading, mocking, or praising. Grundlingh (2018), a linguist, adapted the *speech act* theory of languages, which concerns what utterances *do*, to memes, and gave a taxonomy based on their *illocutionary acts* (Appendix A). For example, stating, predicting, stereotyping, and disputing are some of the illocutionary acts in the taxonomy.

As shown later in Sections 3 and 4, detecting "what the meme does" is of special relevance to CMU work concerning harmful intents. Consequently, existing work in harmful memes has developed multiple fine-grained functional taxonomies for this subset (Sharma et al., 2022a; Banko et al., 2020; Arora et al., 2023; Pramanick et al., 2021a).

2.3 Topics

Memes can also be organized by *topics*, i.e., the semantic themes they are concerned with. Each topic requires a unique set of background knowledge or unique reasoning ability (such as in math or linguistic jokes). Therefore, the topics of the memes may dictate the choice of models to understand them.

There is virtually no fixed set of topical taxonomy for memes, as the internet discusses infinitely many topics. While some topics have existed over the years, such as misogyny (Fersini et al., 2022) and antisemitism (Chandra et al., 2021a), others are *time-sensitive*, based on events happening in real life. For example, recent emerging topics for memes are the US presidential election (Suryawanshi et al., 2020a), COVID-19 (Dimitrov et al., 2021; Pramanick et al., 2021a,b), and the Russia-Ukraine

crisis (Thapa et al., 2024). Such topics require models to be grounded in the latest world news.

3 Tasks

Existing CMU tasks can be broadly divided into three categories, as discussed below.

3.1 Classification

The vast majority of work on CMU has focused on labeling memes with predefined categories[3]. The focus has been on *detecting malicious memes*, such as those that are offensive (Suryawanshi et al., 2020a), trolling (Suryawanshi et al., 2020b), hateful (Kiela et al., 2020), antisemistic (Chandra et al., 2021a), harmful (Pramanick et al., 2021a,b), and misogynous (Fersini et al., 2022). These can be seen as *binary* classification tasks.

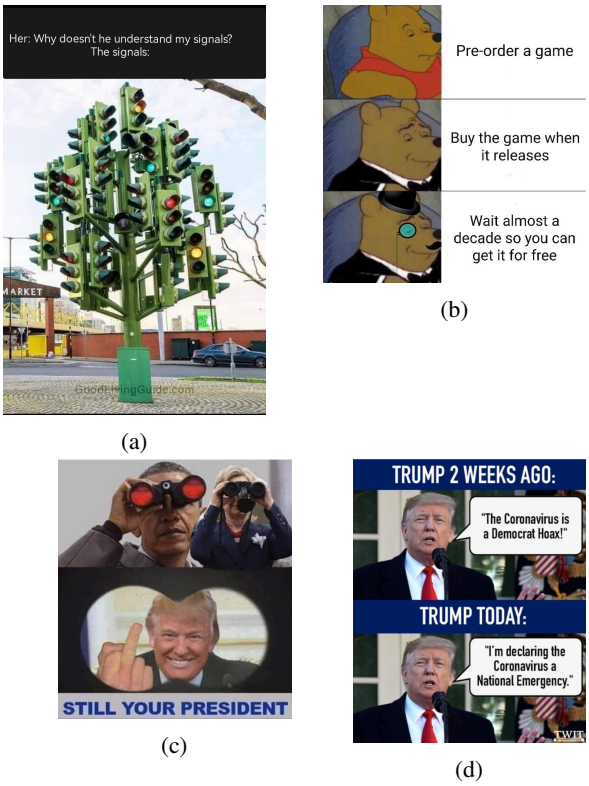
Besides, there are tasks that concern predicting other aspects of memes, such as detecting the persuasion techniques used (Dimitrov et al., 2021), the targets (e.g., religion, race, sex, nationality, or disability) (Mathias et al., 2021; Pramanick et al., 2021a,b), the emotion type (e.g., sarcastic, humorous, motivation, or offensive) (Sharma et al., 2020), the types of figurative language (e.g., allusion, irony/sarcasm, contrast, etc.) (Liu et al., 2022), the roles of people in memes (e.g., hero, villain, or victim) (Sharma et al., 2022b), and meme genres (Dubey et al., 2018; Zannettou et al., 2018; Theisen et al., 2020, 2023)[4]. These are *multi-class* classification problems.

Systems tackling these classification tasks are typically evaluated via accuracy, F1-macro score (i.e., the average F1-score over all classes), and Area Under the ROC Curve (ROC AUC).

3.2 Interpretation

The second category of work involves the relatively new task of *meme interpretation*. This task aims to generate text that captures the final meaning of a meme, which we call the *final message*.

To our knowledge, meme interpretation has only been tackled by Hwang and Shwartz (2023), who refer to the meme interpretation task as *meme captioning* and released the `MemeCap` dataset. An example from `MemeCap` is shown in Figure 2a. This meme was annotated with the caption "*Meme poster is conveying that women wonder why men don't understand their signals when they are overly complicated*".



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

Figure 2: Example memes from (a,b) MemeCap (Hwang and Shwartz, 2023), (c) SemEval-2021-T6 (Dimitrov et al., 2021), and (d) ExHVV (Sharma et al., 2023)

As a text generation task, meme interpretation can be evaluated either manually (i.e., via human evaluation) or automatically using n-gram-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), or semantics-based metrics such as BERTScore (Zhang et al., 2020).

3.3 Explanation

Like the second category, the third category of work also involves generating text, but the focus here is on generating a textual *explanation* of a *label* assigned to the meme, as described below.

Given a harmful meme, an entity in the meme, and a role played by the entity, Sharma (2023) defined the task of generating an explanation of *why* the entity plays the given role in the meme, where the role can be one of "*hero*", "*villain*", and "*victim*". For the meme in Figure 2d, given the entity "*the Democratic Party*" and the role "*victim*", the explanation would be "*The Democratic Party is portrayed as a victim of false allegations*".

[3]There is related work on identifying memes from non-meme images (e.g., Beskow et al. (2020)), which is out of the scope of CMU and is therefore not covered in this paper.

[4]For meme genres, the classes are usually not known in advance.

Hee et al. (2023), on the other hand, addressed the task of explaining the reason for hateful memes.[5] Given a hateful meme and a *general* target (e.g., "race"), the goal is to (1) identify the *specific target group* within the general target (e.g., "*Jews*", "*blacks*") towards which the hate is directed; and then (2) describe *how* the meme poster expressed their hateful feeling towards this specific target group. More specifically, this task involves generating a *reason*, which must follow the pattern: "<*action*> <*target*> <*predicate*>", where <*action*> is either "*use of derogatory term against*" or a single verb, <*target*> is the attacked social target, and <*predicate*> is the hateful implication.

Note that these explanation tasks are different from the meme interpretation task. The explanation tasks can be viewed as *constrained* generation tasks: in Sharma et al.'s task, both the target and the role are given, whereas in Hee et al.'s task, the general target is given. In contrast, such constraints are not present in the meme interpretation task. As an example, consider the meme in Figure 2d again. The final message that we would have produced for this meme (as the output of meme interpretation) is "*The meme poster makes fun of Trump for the change in his recognition of the severity of the Coronavirus*", which is very different from the explanation being generated when the target is constrained to be "*the Democratic Party*".

3.4 Challenges to Meme Understanding

To better understand the difficulty of CMU, below we discuss the unique challenges posed by the task at a high level.

Meme-specific Knowledge On top of the challenges in retrieving knowledge about the physical and cultural world as with any commonsense reasoning task, memes further require a broad understanding of the *meme culture*. According to Milner (2012), meme comprehension requires *subcultural literacy* – the "insider's knowledge" maintained by various internet subcommunities. To enable meme understanding, it is important for computer systems to automatically access this knowledge.

A typical type of internet-culture literacy required is the ability to make use of the form of a meme to infer its meaning. In the *Macros* type (Section 2.1), the audience is assumed to know how the *base template* works with the filled text to create the final meaning. Meanwhile, *Stack Images* comprise multiple images in a meme. Those images can follow a chronological order (Figure 2d), respond to each other (Figure 2c), or form a predefined *template* (Figure 2b). One must understand the meaning of these forms to combine separate elements in a meme and infer the final meaning.

Temporal Context A meme is implicitly assumed (by its author) to be read in the context of the date on it was posted. Hence, to correctly understand a meme, a system must have the correct context as the internet audience at that time. Consider the meme in Figure 2c illustrating Trump saying "Still your president". Given that Trump was not re-elected as the US president in the 2020 presidential election, if we start with the wrong knowledge that the meme was posted in 2024, the meme would suggest that Trump is still the president today, which is contradictory to current knowledge. Only when knowing the temporal context that the meme was posted around 2020 could one infer the correct meaning of the meme, which is "*Trump is still America's president (in 2020)*".

This temporal-contextual property of memes warrants a sense of time when models retrieve their knowledge. If the system is trained with data *after* the post date of the meme, it has to "think in the past", i.e., not using the information after the post date. A harder case is when a system was trained only on data *before* the post date of the meme, such as a model that monitors hate speech in social networks in real time. This system needs to acquire knowledge as up-to-date as current internet users. If the system cannot catch up quickly enough, bad consequences may already happen — for Figure 2c, it means that hatred has already been spurred on the internet and reputations damaged. Therefore, as memes are imitated and circulated quickly and out-of-context on the internet, the challenge lies in maintaining post dates, staying up-to-date, as well as thinking in the past.

Subjectivity in Interpretation For the meme interpretation task, annotators are *not* asked to write down the message that the meme author tries to convey, since in reality there is no way to verify the author's intent. Rather, they are asked to write down the message that *they believe* the author tries to convey. Hence, annotating memes with messages for the meme interpretation task is inherently subjective since readers with different backgrounds

[5] While being harmful means "having the potential to cause harm to individual, groups, or society", being hateful is a specific type of harm that attacks a group by their characteristic (e.g., race, religion, gender).

may understand a meme differently. For example, a meme can be offensive to one reader, but may not be so to another. As a consequence, it is possible for a meme to have more than one interpretation, as admitted in Sharma et al. (2023).

The inherent subjectivity, however, does not imply that all messages that annotators come up with should be accepted as gold-standard messages. In particular, there is still a notion of correctness that can be defined. For instance, an annotator may have misrecognized or simply ignored some crucial visual cues in the input or made some unwarranted assumptions in their reasoning process, resulting in erroneous messages. A challenge, then, involves designing an annotation mechanism that can facilitate the identification of such messages.

Furthermore, since multiple messages may be considered correct for a given meme, a meme interpretation model can output all plausible interpretations. Realistically, however, we believe what matters the most is whether the model can output the most popular messages (i.e., the messages that most people perceive to be what the author tries to convey). Hence, the challenge is now to acquire a model with this notion of popularity.

Interpretable Models for Meme Interpretation The users of a meme interpretation model are supposed to be the general public on social media platforms. They have the right to question the correctness of the output, such as when the model flags a meme to be harmful and should be removed. How can we increase a user's confidence in the model's output in such cases?

One plausible solution is *interpretability*. If the model can explain the message(s) it outputs for a meme, a user can inspect the explanation and determine whether the message(s) can be trusted. While short, often one-sentence, explanations are expected in many interpretability tasks (Sharma et al., 2023; Hee et al., 2023; Lin et al., 2024), a convincing explanation for meme interpretation may need to be more detailed. Ideally, it should mimic the human reasoning process, which can be seen as a multi-step derivation process that combines the textual and visual cues with the relevant hidden information (background knowledge and beliefs) to arrive at the final message. Considering that explaining the output of neural models is still an open problem even for generating short explanations (Räuker et al., 2023), designing models that can output explanations as detailed as a human reasoning process is conceivably very challenging.

4 Datasets

In this section, we present existing CMU datasets.

4.1 Overview

Table 1 categorizes 24 commonly-used datasets by the type of tasks for which they were created (as defined in Section 3).

Classification Nearly all datasets were designed for classification tasks (21/24). Among those, 16 are about detecting or classifying malicious memes (see the first group of rows in Table 1). The variations of malignity are hate, harm, offensiveness, trolling, abuse, antisemitism, misogyny, and aggression. In addition, five classification datasets (the second group of rows in Table 1) contain labels for aspects of memes that are *orthogonal* to harmful attacks, including sentiments, emotions, types of figurative language, semantic roles of entities, and persuasion techniques.

Interpretation So far, `MemeCap` (Hwang and Shwartz, 2023) is the only dataset created for meme interpretation. To construct `MemeCap`, the authors first scraped the images from Reddit and made sure that they were non-offensive, non-sexual, and that the text and the image complement each other. Then, for each meme, they manually annotated the meme captions, the *literal captions* (i.e., the caption of the image excluding the text) and the *visual metaphors* (i.e., the associations between the entities in the meme and its actual target).

Explanation `HatReD` and `ExHVV`, the two datasets that contain ground-truth labels for explanations, were both created by adding a new layer of annotation for explanations on existing datasets that already have categorical labels. For `HatReD`, the possible labels are "hateful" and "non-hateful". For `ExHVV`, the labels are roles (i.e., "hero", "villain", or "victim"). Multiple explanations per label were allowed.

4.2 Discussion

We discuss the issues with existing datasets.

4.2.1 Forms Overlooked

According to our taxonomy for memes, current datasets have *not* covered all the possible meme types.

Regarding forms, there are some studies where only one form of memes is considered. For example, in the creation of the `HatefulMemes` dataset

Dataset and/or Publication	Task	Objective	# Memes	Lang.	Method	License
HatefulMemes (Kiela et al., 2020)	2C	Hate	10,000	E	**Synthesis**	Custom
MUTE (Hossain et al., 2022b)	2C	Hate	4,158	E+**Be**	Scrape	MIT
MMHS150K (Gomez et al., 2019)	2C	Hate	150,000	E	Scrape	Custom
Sabat et al. (2019)	2C	Hate	5,020	E	Scrape	CC0
CrisisHateMM (Thapa et al., 2024)	NC	Hate & Target	4,486	E	Scrape	MIT
WOAH-5 (Mathias et al., 2021)	NC	Hate Type & Target	10,000	E	**Inherit**	Apache-2.0
HarMeme (Pramanick et al., 2021a)	2C, NC	Harm & Target	3,544	E	Scrape	BSD
HARM-C&P (Pramanick et al., 2021b)	2C, NC	Harm & Target	7,096	E	**Inherit**	MIT
Giri et al. (2021)	NC	Offensiveness	6,992	E	Scrape	Unavailable
Shang et al. (2021b)	2C	Offensiveness	3,059	E	Scrape	Unavailable
MultiOFF (Suryawanshi et al., 2020a)	2C	Offensiveness	743	E	Scrape	None
TamilMemes (Suryawanshi et al., 2020b)	2C	Trolling	2,969	**T**	Scrape	GNU-3.0
BanglaAbuse (Das and Mukherjee, 2023)	2C	Abuse	4,043	**Be**	Scrape	MIT
Jewtocracy (Chandra et al., 2021a)	2C, NC	Antisemitism	6,611	E	Scrape	Unavailable
MAMI (Fersini et al., 2022)	2C, NC	Misogyny	11,000	E	Scrape	Apache-2.0
MIMOSA (Ahsan et al., 2024)	NC	Agression Target	4,848	**Be**	Scrape	MIT
Memotion (Sharma et al., 2020)	NC	Emotion	10,000	E	Scrape	MIT
FigMemes (Liu et al., 2022)	NC	Figurative Lang.	5,141	E	Scrape	None
HVVMemes (Sharma et al., 2022b)	NC	Role of Entities	7,000	E	**Inherit**	None
MemoSen (Hossain et al., 2022a)	NC	Sentiment	4,417	**Be**	Scrape	Custom
SemEval-2021-T6 (Dimitrov et al., 2021)	NC	Persuasion Tech.	950	E	Scrape	None
HatReD (Hee et al., 2023)	E	Hate	3,304	E	**Inherit**	Custom
ExHVV (Sharma et al., 2023)	E	Role of Entities	4,680	E	**Inherit**	CC0-1.0
MemeCap (Hwang and Shwartz, 2023)	I	Meme Captioning	6,387	E	Scrape	GPL-3.0

Table 1: **Existing Datasets on Computational Meme Understanding.** Abbreviations: for **Task** – Binary classification (2C), Multi-class classification (NC), Explanation (E), and Interpretation (I); for **Method** – "Inherit" means the memes were from another dataset.; for **Lang.** (Languages): English (E), Bengali (Be), T (Tamil).

and the meme analysis study of Zhou et al. (2023), only the Macros form is considered. According to Kirk et al. (2021), because memes from the `HatefulMemes` dataset do not cover other types of memes such as *Screenshots* (of conversations) or plain text, models that are trained on it suffer when dealing with memes "in the wild". Therefore, knowing that Macros are only one of many forms, it is worth noting that the scope of such studies is limited. For other datasets where memes were scraped, we are not aware of any deliberate control of the forms of memes. Therefore, it remains unclear if the datasets cover all the types of memes.

For the other two dimensions of our taxonomy, existing dataset authors have monitored them rather satisfactorily. Topical distribution was either automatically kept track of via the search keywords or calculated via manual inspection on a sample (Hwang and Shwartz, 2023). The functions of the memes, on the other hand, are actual labels of the task, which are kept track of by default.

4.2.2 Annotation Quality

Classification datasets: Inter-annotator agreement is not always reported We examined five datasets that were organized as shared tasks, namely `SemEval-2021-T6`, `Memotion`, `WOAH-5`, `MAMI`, and `HatefulMemes`. The first two did not report the inter-annotator agreement, and while the remaining three did, `MAMI` only has a Kappa score of 0.33, which implies "fair" agreement (Viera and Garrett, 2005). Given the subjectivity inherent in meme understanding, it is important to report agreement measures to gauge data quality and form a realistic expectation for model performances.

Interpretation dataset: No review of annotations In `MemeCap`, all annotations were produced by Mechanical Turkers from geographically diverse backgrounds, which is good in eliciting multiple opinions when collecting captions. The procedure has started to address the challenge of multiple messages (Section 3.4) via collecting multiple annotations per instance. However, it appears that no review of the annotations was conducted. This raises questions about the quality of the data.

To control annotation quality, the authors of `HatReD` and `ExHVV` (Hee et al., 2023; Sharma et al., 2023) organized multiple rounds of training for the annotators. The manually annotated explanations were scored by multiple human judges on multiple aspects. This approach of quality control is also known as COLLECT-AND-JUDGE (Wiegreffe and Marasovic, 2021), which is a good practice. `HatReD` even went one step further and re-

ported the "inter-judge" agreement of the judges, which helped increase the reliability of the scoring process. However, COLLECT-AND-JUDGE still risks having the judges biased toward giving higher scores. If the bias is shared between judges, the reported quality can still be *artificially* high.

4.2.3 Temporal Context

Section 3.4 posed the challenge of recognizing the temporal context of a meme. One can start tackling this challenge by collecting the posted timestamps of memes. However, none of the datasets we found recorded this information. Some datasets specify the date range during which the memes were collected, which is a good first step.

5 Models

In this section, we present an overview of the models that have been built for the three CMU tasks.

5.1 Classification Models

Approaches The majority of meme classification systems follow a consistent recipe. Given a meme, a system first extracts important features such as the filled text and the properties of the entities in the image (e.g., races and genders) using off-the-shelf models or API services such as Google Cloud Vision API, Easy-OCR[6], and FairFace (Kärkkäinen and Joo, 2019). Next, one may encode the visual and textual information of the meme into an embedding space using one of the vision encoders, such as ResNet (He et al., 2015), VilBERT (Lu et al., 2019), ViT (Dosovitskiy et al., 2021), CLIP (Radford et al., 2021), and Perceptual Hashing (Monga and Evans, 2006), SURF (Bay et al., 2008), and one of the language encoders, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2023), and Llama 2 (Touvron et al., 2023). The two modalities will then be *aligned* using concatenation or techniques like Cross-Attention (Lin et al., 2021). Finally, the vector representation of all aspects of the meme will be fed into a classification head such as a Feedforward Neural Network to generate the label.

Alternatively, one may reduce the multimodal problem into a text classification task by first generating a textual description of the image, then take the image description, the OCR text, and other features as the model inputs (Cao et al., 2023, 2022).

Recently, many newly-released vision-language models (VLMs), which are built by fusing a language model and a vision encoder, have shown strong performances in meme classification after fine-tuning. While many of these models are proprietary, such as Flamingo (Alayrac et al., 2022), PaLI (Chen et al., 2023), and GPT4 (OpenAI et al., 2024), there are also open-sourced versions such as Llava (Liu et al., 2023) and OpenFlamingo (Awadalla et al., 2023).

Performances There is a spectrum of model performance across benchmarks. Some models have achieved accuracy levels above 90%, such as PaLI-X-VPD (Hu et al., 2024) on `HatefulMemes` (binary classification) and a combination of CLIP (Radford et al., 2021), LASER (Artetxe and Schwenk, 2019), and LaBSE (Feng et al., 2022) on `WOAH5` (5 and 7 classes). However, some benchmarks are particularly challenging, such as `SemEval-2021-T6`, where the best model only achieved an F1 score of 0.58 over 22 classes. These results show that models still have a lot of room for improvement in tasks that are as complex as `SemEval-2021-T6`.[7]

An overview of the state-of-the-art models for meme classification tasks and their performances can be found in Table 2 (Appendix B).

5.2 Explanation Models

Approaches Models for meme explanation extend those for classification by replacing the classification head by a language decoder to generate text (Hee et al., 2023). LUMEN, the system proposed with `ExHVV`, was built via *joint learning* for classification and explanation tasks.

Performances The explanation models mostly score low in human evaluation. Unfortunately, LUMEN's authors did not report human evaluation results. Therefore, we do not have qualitative insights into the challenges faced by this system. Meanwhile, for `HatReD`, the best systems were shown to score under 70% w.r.t. correctness. In their human evaluation, Hee et al. (2023) showed that the model's performance was hurt by unreliable visual information extractors. Furthermore, hallucinations emerge in the model's results. The authors recommend that future efforts be expended on "using retrieval augmentation to incorporate explicit knowledge", which suggests that their system is also struggling with leveraging meme-specific and topic-specific knowledge.

Details on the performances of the state-of-the-art meme explanation and interpretation models

[6] https://github.com/JaidedAI/EasyOCR

[7] Common errors are discussed in Appendix C.

can be found in Table 3 (Appendix D).

5.3 Interpretation Models

For MemeCap, the only interpretation dataset released so far, the authors experimented with the open-source versions of the state-of-the-art vision-language models. Results paint a similar picture as in the explanation task, where models still struggle to infer the correct meaning of the memes. Hwang and Shwartz (2023) showed that the models' errors usually arose from (1) the failure to attend to important visual elements and (2) the lack of sufficient background knowledge.

6 Ethical Considerations

Memes are an effective means of communication. Therefore, any technology that harnesses the power of memes also carries the risk of being misused and other negative side effects to humans. This section highlights two major ethical considerations related to the development of CMU technologies.

Should we expose annotators to sexual and hateful content? The main motivation for annotating memes so far has been to automatically process malicious contents such as manipulation, including hateful memes (Kiela et al., 2020) and harmful memes (Pramanick et al., 2021a). However, exposing annotators to such content may negatively affect their mental health. As such, researchers working on meme annotation should impose age restrictions on annotators and have appropriate pre-annotation screening to ensure the candidate annotators understand the potential harm of the data before joining the annotation team.

In a crowdsourcing setting, however, some people may choose to become annotators because the cost of forfeiting the opportunity is too high, thus accepting the risks to their mental health. Therefore, if the data is harmful, one should hire annotators instead of perform crowdsourcing so that they can monitor the mental health of the participants.

Should we create datasets and models on harmful memes? Recall that one of our goals is to help identify hateful memes or display warning messages. However, there is a risk that the datasets and models will be misused to generate hateful messages or bypass hate detection models (i.e., adversarial attacks). Therefore, the public release of such data and models should be handled with care, ideally after proper consultation from social science researchers. For datasets, one may adopt the licensing approach similar to that employed by the authors of the Hateful Memes dataset (Kiela et al., 2020). For models, *red teaming* (Ganguli et al., 2022) should be conducted rigorously before deploying them to the real world.

7 Concluding Remarks

We conclude this survey by enumerating promising avenues for future research in CMU.

Richer Annotations for More Robust Models For CMU models, particularly those that concern meme interpretation, to be robustly deployed, they should be more powerful than they currently are. Not only should they be interpretable, but they should be equipped with the ability to rank plausible messages based on their popularity. In particular, for the task of meme interpretation, building interpretable models involves mimicking the human reasoning process, which is extremely challenging.

Given the complexity of these learning tasks, one could consider employing supervised approaches, at least in the initial stage of the learning process. To facilitate such approaches, researchers should study how to best represent such reasoning processes (from the perspective of annotation) and start collecting training data for this new task.

Improving Annotation Procedures with VLMs In meme explanation, Lin et al. (2024) showed that VLMs are good and even surpassed humans in certain metrics. Therefore, there is a clear potential to reduce human annotation effort for CMU tasks by exploiting VLMs in the process. For example, we can have a VLM produce an initial draft before asking human annotators to review it to ensure that it free of hallucinations. However, editing text may be more time-consuming than writing text from scratch. Therefore, further investigations on the benefits and drawbacks of incorporating VLMs are warranted.

Next Level of Visual Reasoning Current models often miss the deciding visual elements that are important to the meaning of a meme (Hwang and Shwartz, 2023). For example, failing to recognize the crucial demographic information of a person in a given meme makes it difficult for models to follow the correct line of reasoning (Hee et al., 2023). Here, the reason for missing details was that the systems relied on a *task-agnostic* image captioner that was used to translate *everything* on the image to text. When the image caption misses an important detail, the whole pipeline fails.

How can we teach models to attend to the "right" details? A meme usually has many semiotic resources (Grundlingh, 2018), but not all are important in conveying the message(s). Humans usually talk about interpretations by pointing to the visual elements and explaining them in words. Hence, demonstrations of what visual elements to attend are likely to be helpful for models. To that end, we encourage the construction of datasets to contain a textual explanation of the human's *reasoning process* in understanding a meme, in which the visual details for meme understanding are mentioned explicitly. The resulting data source will be useful for training models in directing visual attention.

Active Knowledge Acquisition For real-world deployment, models need to acquire knowledge in the wild because the context in which a meme is created and shared can change quickly. At least two types of knowledge are crucial for CMU.

Meme cultures: The knowledge here refers to the understanding of what the meme templates mean. This knowledge can perhaps be accessed by leveraging the internet databases of memes. For example, Know Your Meme[8] is "the world's largest internet culture authority", where one can access documentation about memes and other internet phenomena. Hence, a model trained on this database can acquire knowledge about meme cultures. Model developers may periodically pull data from this database to update a model's knowledge.

Topic-specific background knowledge: For instance, in mocking memes, rather than being explicitly mentioned, the mocked target may be implied from some characteristics mentioned in the meme. Here the implicit knowledge is the association between such characteristics and the target. How to acquire such kind of knowledge remains an open question. One possibility would be to examine how Retrieval Augmentation (Gao et al., 2024) can be leveraged in this multimodal setting.

Connection to Pragmatics Memes are hard to understand largely because they require contextual information. Pragmatics, including the processing of presuppositions and deixis, as well as social-context grounding can provide insights into identifying high-level features to improve performance on CMU tasks. For example, consider the meme in Figure 2c and how pragmatics can help improve models' performance. The text "Still your president" has the presupposition that Trump was previously the US president, which should be verified as background knowledge. Additionally, the image of Obama and Clinton using binoculars to watch Trump presupposes that Obama and Clinton, who represent the Democratic party, are stalking Trump. This starts to reveal the opinion of the meme. Furthermore, the deixis "your" in the text, when being resolved as being used by Trump to reference Obama and Clinton, can help highlight the tension between the two parties. Finally, by grounding the social context when the meme was happening, which includes the fact that the Democrats are trying to vote Trump out of office, one can be more confident that the meme is indeed "praising Trump and mocking the opposite party". Hence, by recognizing presuppositions, deixes, and social context, one can make them available as input to CMU systems, which has the potential to improve model performance.

Towards Processing Animated and Video Memes Memes do not only exist in static images. GIFs and short videos are even more widespread. For instance, GIPHY, the world's biggest GIF site, saw more than one billion searches for GIFs every day.[9] Meanwhile, Youtube Shorts, one of the leading short-video platforms, saw 50 billion daily views recently.[10] Automatically understanding memetic contents in these formats can unlock a huge source of information for the study of online communications. However, GIFs and short videos present notable challenges as they are technically composed of many frames of images, exhibiting complex relationships between video frames. Therefore, it would be fruitful to approach CMU for GIFs and short videos.

Meme Generation Meme generation is an important next step beyond meme understanding, for at least three reasons. First, meme generation could be used to measure a model's understanding of how memes work (via the generated memes). Second, meme generation, like humor generation, has huge applications in humanizing computer interfaces by making them more humorous, friendly and trustworthy (Hempelmann, 2008). Finally, technologies for automatically generating captivating online content could have a big impact on digital marketing and other fields related to online communications.

[8] https://knowyourmeme.com/

[9] https://www.ads.giphy.com/

[10] https://techcrunch.com/2023/02/03/google-says-youtube-shorts-has-crossed-50-billion-daily-views/

Limitations

Owing to space limitations, we have only been able to provide a high-level overview of the models used for the various CMU tasks. In particular, since the survey does not go into the details of individual models, we did not highlight the strengths and weaknesses of each model. Rather, we discuss the strengths and weaknesses of existing models in a collective fashion and refer the reader to the original papers if they are interested.

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500, St. Julian's, Malta. Association for Computational Linguistics.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. 2023. Detecting Harmful Content On Online Platforms: What Platforms Need Vs. Where Research Efforts Go. ArXiv:2103.00153 [cs].

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. ArXiv:2308.01390 [cs.CV].

Kent Bach and Robert M. Harnish. 1984. *Linguistic communication and speech acts*, 1.ed., 2. print edition. MIT Press, Cambridge, Mass.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359. Similarity Matching in Computer Vision and Multimedia.

David M. Beskow, Sumeet Kumar, and Kathleen M. Carley. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, 57(2):102170.

Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang. 2023. ProCap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection. ArXiv:2308.08088 [cs].

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021a. "Subverting the Jewtocracy": Online Antisemitism Detection Using Multimodal Deep Learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, pages 148–157, New York, NY, USA. Association for Computing Machinery.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021b. "Subverting the Jewtocracy": Online Antisemitism Detection Using Multimodal Deep Learning. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, pages 148–157, New York, NY, USA. Association for Computing Machinery.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. Pali: A jointly-scaled multilingual language-image model. ArXiv:2209.06794 [cs].

Mithun Das and Animesh Mukherjee. 2023. BanglaAbuseMeme: A dataset for Bengali abusive meme classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15498–15512, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv:2010.11929 [cs].

Abhimanyu Dubey, Esteban Moro, Manuel Cebrian, and Iyad Rahwan. 2018. MemeSequencer: Sparse Matching for Embedding Image Macros. ArXiv:1802.04936 [cs].

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, and Li Chen. 2021. Alpha at SemEval-2021 task 6: Transformer based propaganda classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 99–104, Online. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. ArXiv:2209.07858 [cs].

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv:2312.10997 [cs].

Roushan Kumar Giri, Subhash Chandra Gupta, and Umesh Kumar Gupta. 2021. An approach to detect offence in Memes using Natural Language Processing(NLP) and Deep learning. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5, Coimbatore, India. IEEE.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring Hate Speech Detection in Multimodal Publications. ArXiv:1910.03814 [cs].

Lezandra Grundlingh. 2018. Memes as speech acts. *Social Semiotics*, 28(2):147–168.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. ArXiv:1512.03385 [cs].

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the Underlying Meaning of Multimodal Hateful Memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5995–6003, Macau, SAR China. International Joint Conferences on Artificial Intelligence Organization.

Christian F Hempelmann. 2008. Computational humor: Beyond the pun? *The Primer of Humor Research. Humor Research*, 8:333–360.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022a. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022b. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024. Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models. ArXiv:2312.03052 [cs].

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Saurav Joshi, Filip Ilievski, and Luca Luceri. 2024. Contextualizing internet memes across social media platforms. ArXiv:2311.11157 [cs].

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, volume 33, pages 2611–2624, Virtual. Curran Associates, Inc.

Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 26–35, Online. Association for Computational Linguistics.

Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. ArXiv:1908.04913 [cs].

Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling Hate in Online Memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 5138–5147, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hezheng Lin, Xing Cheng, Xiangyu Wu, Fan Yang, Dong Shen, Zhongyuan Wang, Qing Song, and Wei Yuan. 2021. CAT: Cross Attention in Vision Transformer. ArXiv:2106.05786 [cs].

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models. ArXiv:2401.13298 [cs].

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*, New Orleans, Louisiana.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. ArXiv:1908.02265 [cs].

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A Survey on Computational Propaganda Detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4826–4832, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online. Association for Computational Linguistics.

Ryan M Milner. 2012. *The world made meme: Discourse and identity in participatory media*. Ph.D. thesis, University of Kansas.

V. Monga and B.L. Evans. 2006. Perceptual image hashing via feature points: Performance evaluation and tradeoffs. *IEEE Transactions on Image Processing*, 15(11):3452–3465.

Vincent Ng and Shengjie Li. 2023. Multimodal Propaganda Processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15368–15375.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. ArXiv:2303.08774 [cs].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy

Chakraborty. 2021b. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. ArXiv:2103.00020 [cs].

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs, stat].

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. ArXiv:2207.13243 [cs].

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation. ArXiv:1910.02334 [cs].

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Lanyu Shang, Christina Youn, Yuheng Zha, Yang Zhang, and Dong Wang. 2021a. KnowMeme: A Knowledge-enriched Graph Neural Network Solution to Offensive Meme Detection. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 186–195.

Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. 2021b. AOMD: An Analogy-aware Approach to Offensive Meme Detection on Social Media. ArXiv:2106.11229 [cs].

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Gauri Sharma. 2023. Discovering safety issues in text-to-image models: Insights from adversarial nibbler challenge. In *Proceedings of the ART of Safety: Workshop on Adversarial testing and Red-Teaming for generative AI*, pages 43–48, Bali, Indonesia. Association for Computational Linguistics.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What Do You MEME? Generating Explanations for Visual Semantic Role Labelling in Memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022a. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5597–5606. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022b. Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020b. A dataset for troll classification of TamilMemes. In *Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Jafri, Hariram Veeramani, Raghav Jain, Sandesh Jain, Francielle Vargas, Ali Hürriyetoğlu, and Usman Naseem. 2024. Extended multimodal hate speech event detection during Russia-Ukraine crisis - shared task at CASE 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 221–228, St. Julians, Malta. Association for Computational Linguistics.

William Theisen, Joel Brogan, Pamela Bilo Thomas, Daniel Moreira, Pascal Phoa, Tim Weninger, and Walter Scheirer. 2020. Automatic discovery of political meme genres with diverse appearances. ArXiv:2001.06122 [cs].

William Theisen, Daniel Gonzalez Cedre, Zachariah Carmichael, Daniel Moreira, Tim Weninger, and Walter Scheirer. 2023. Motif Mining: Finding and Summarizing Remixed Image Content. In

2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1319–1328, Waikoloa, HI, USA. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Anthony J. Viera and Joanne M. Garrett. 2005. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. ArXiv:1805.12512 [cs].

Jing Zhang and Yujin Wang. 2022. SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, Seattle, United States. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. ArXiv:1904.09675 [cs].

Naitian Zhou, David Jurgens, and David Bamman. 2023. Social Meme-ing: Measuring Linguistic Variation in Memes. ArXiv:2311.09130 [cs].

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219, Online. Association for Computational Linguistics.

A Illocutionary Acts of Memes

In pragmatics (the study of how language means differently in different contexts), the speech acts theory identifies three main types of action that a spoken utterance can do: locution (the production of sounds and words), illocution (performing one of the functions of language), and perlocution (the effects resulted from saying something). When studying memes, Grundlingh (2018) has argued that memes also demonstrated similar illocutionary acts as spoken utterances and proposed a taxonomy for linguistic functions of memes (Figure 3), adapted from the communicative illocutionary acts identified by Bach and Harnish (1984). This taxonomy can be seen as a theory-based vocabulary for meme functions.

B State-of-the-Art Models for Meme Classification Tasks

Table 2 details the most performant models for memes classification tasks. Details about these datasets have been shown in Table 1. Many state-of-the-art models are proposed by the dataset authors themselves. While some benchmarks received a lot of attention and quickly became saturated such as Hateful Memes and WOAH 5, many other benchmarks do not see many improvements over the years.

C Common Errors in Meme Classifiers

Chandra et al. (2021b), when analyzing the errors in their antisemitism detector, showed that the lack of context of the memes caused the models to misclassify. Additionally, Cao et al. (2023) and Pramanick et al. (2021a) showed that biased data during training (e.g., most images with Muslims are flagged as hateful) caused the models to be biased towards classifying certain topics as hateful regardless of the actual content. Other causes include models failing to perform complex reasoning processes on the text (Chandra et al., 2021b) or failing to attend to the important visual information (Pramanick et al., 2021b).

D State-of-the-Art Models for Meme Explanation and Interpretation

Table 3 details the best-performing models for the meme explanation and interpretation tasks. Except LUMEN, all models are pretrained models prompted or fine-tuned on the corresponding tasks.

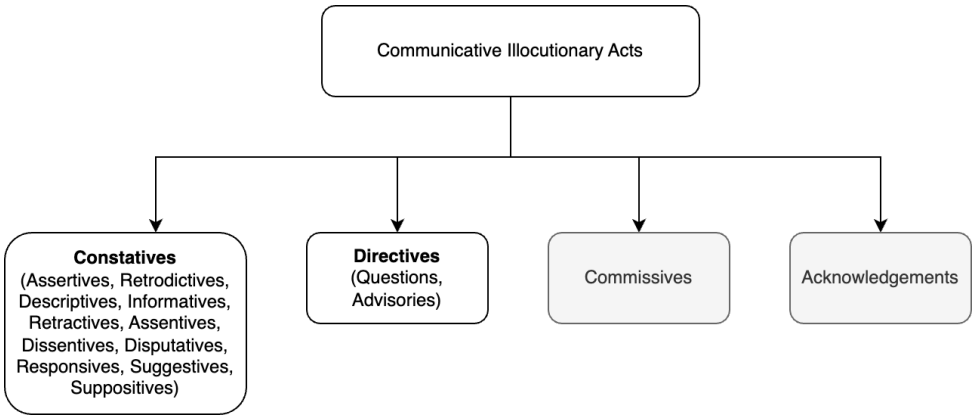


Figure 3: Illocutionary acts of memes, adapted from Grundlingh (2018). The text in parentheses represents subtypes. Commissives and Acknowledgements (in gray) are illocutionary acts from speech acts theory that do not apply to memes.

Publication of state-of-the-art models	Dataset	Task	Acc	AUC	F1
Hu et al. (2024)	Hateful Memes (Kiela et al., 2020)	B	.90	.81	
Zia et al. (2021)	WOAH5 (Mathias et al., 2021)	N T.		.96	
Mathias et al. (2021)		N T.			.91
Zia et al. (2021)		N A.		.97	
Mathias et al. (2021)		N A.			.91
Cao et al. (2023)	MAMI (Fersini et al., 2022)	B	.74	.84	
Zhang and Wang (2022)		B			.83
Zhang and Wang (2022)		N T.			.73
Cao et al. (2023)	HarMeme (Pramanick et al., 2021a)	B	.91		
Pramanick et al. (2021a)		N L.	.76		.54
Pramanick et al. (2021a)		N T.	.76		.66
Lin et al. (2024)	HARM-C (Pramanick et al., 2021b)	B	.87		.86
Pramanick et al. (2021b)		N L.	.77		.55
Pramanick et al. (2021b)		N T.	.78		.70
Lin et al. (2024)	HARM-P (Pramanick et al., 2021b)	B	.91		.91
Pramanick et al. (2021b)		N L.	.87		.67
Pramanick et al. (2021b)		N T.	.79		.69
Chandra et al. (2021b)	Jewtocracy (Chandra et al., 2021a)	B Tw	.72		
Chandra et al. (2021b)		B G.	.91		
Chandra et al. (2021b)		N Tw	.68		
Chandra et al. (2021b)		N G.	.67		
Lee et al. (2021)	MultiOFF (Suryawanshi et al., 2020a)	B			.65
Suryawanshi et al. (2020b)	TamilMemes (Suryawanshi et al., 2020b)	B			.52
Gomez et al. (2019)	MMHS150K (Gomez et al., 2019)	B	.68	.73	.70
Sabat et al. (2019)	Sabat et al. (2019)	B	.83		
Giri et al. (2021)	Giri et al. (2021)	B	.71		
Giri et al. (2021)		N	.99		
Shang et al. (2021a)	Shang et al. (2021a)	B R.	.73		.49
Shang et al. (2021a)		B G.	.70		.55
Feng et al. (2021)	SemEval-2021-T6 (Dimitrov et al., 2021)	N 3			.58
Sharma et al. (2020)	Memotion (Sharma et al., 2020)	N St.			.35
Sharma et al. (2020)		N H.			.52
Sharma et al. (2020)		N Sm			.32

Table 2: **State-of-the-art models on Meme Classification.** B: Binary classification. N: Multiclass classification. L: Level, T: Target, A: Attack type, G: Gab, Tw: Twitter, R: Reddit, St: Sentiment, H: Humor, Sm: Semantic

Dataset	Model	Automatic Eval.			Human Eval.	
		BLEU	ROUGE-L	BERT	Fluent	Correct
HatReD	Text-only: RoBERTa-base	0.177	0.389	**0.480**	**0.975**	0.544
	Text-only: T5-Large	**0.190**	**0.392**	0.479	0.926	**0.622**
ExHVV	LUMEN	**0.313**	**0.294**	**0.902**		
MemeCap	Open-Flamingo few-shot	0.267	**0.435**	0.739	0.933	**0.361**
		0.270	**0.435**	0.743		
	Llama fewshot	0.266	0.434	**0.747**	**0.967**	**0.361**

Table 3: Best performing models for meme explanation (first two datasets) and interpretation (last dataset). The scores were taken from the respective papers and scaled to the range [0, 1]. The best results for each dataset are **boldfaced**.

No models can score higher than 65% on Correctness in human evaluation.