MEME*X*: Detecting Explanatory Evidence for Memes via Knowledge-Enriched Contextualization

Anonymous ACL submission

Abstract

Memes are a powerful tool for communication over social media. Their affinity for evolving across politics, history, and sociocultural phenomena renders them an ideal vehicle for communication. To comprehend the subtle message conveyed within a meme, one must understand the relevant background that facilitates its holistic assimilation. Besides digital archiving of memes and their metadata by a few websites like `knowyourmeme.com`, currently, there is no efficient way to deduce a meme's context dynamically. In this work, we propose a novel task, MEME*X* – given a meme and a related document, the aim is to mine the context that succinctly explains the background of the meme. At first, we develop `MCC` (Meme Context Corpus), a novel dataset for MEME*X*. Further, to benchmark `MCC`, we propose `MIME` (MultImodal Meme Explainer), a multimodal neural framework that uses external knowledge-enriched meme representation and a multi-level approach to capture the cross-modal semantic dependencies between the meme and the context. `MIME` surpasses several unimodal and multimodal systems and yields an absolute improvement of $\approx$ 4% F1-score over the best baseline. Lastly, we conduct detailed analyses of `MIME`'s performance, highlighting the aspects that could lead to optimal modeling of cross-modal contextual associations.

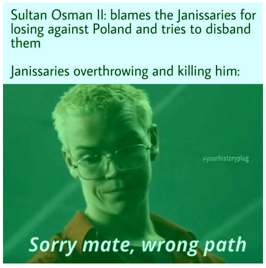
	Meme Source: Reddit - `r/historymemes` Context Source: Wikipedia **The result was a palace uprising by the Janissaries, who promptly imprisoned the young sultan in Yedikule Fortress in Istanbul, where Osman II was strangled to death.** After Osman's death, his ear was cut off and represented to Halime Sultan and Sultan Mustafa I to confirm his death and Mustafa would no longer need to fear his nephew. It was the first time in the Ottoman history that a Sultan was executed by the Janissaries.

Table 1: MEME*X* – given a meme and a relevant context, the aim is to identify the *evidence* in the context that can succinctly explain the background of the meme, depicted above via emboldened and highlighted excerpt.

1 Introduction

Social media has evolved to be a mainstream communication medium for the masses and has thereby redefined how we interact within society. The information shared on social media has diverse forms, like text, audio, and visual messages, or their combinations thereof. A meme is a typical example of such social media artifact that is usually disseminated with the flair of sarcasm or humor. While memes facilitate convenient means for propagating complex social, cultural, or political ideas via visual-linguistic semiotics, they often abstract away the contextual details that would typically be necessary for the uninitiated. Such contextual knowledge is critical for human understanding and computational analysis alike. We aim to address this requirement by contemplating solutions that facilitate the automated derivation of contextual *evidence* towards making memes more accessible. To this end, we formulate a novel task – **MEME*X***, which, given a meme and a related context, aims to detect the sentences from within the context that can potentially explain the meme. Table 1 visually explains MEME*X*.

Table 1 primarily showcases a meme's figure (left) and an excerpt from the related context (right). This meme is about the revenge killing of an *Ottoman Sultan*, by the *Janissaries* (infantry units), in reaction to their disbanding, by the Sultan. The first line conveys the supporting evidence for the meme from the related context, emboldened and highlighted in Table 1. The aim is to model the required cross-modal association that facilitates the detection of such supporting pieces of evidence from a given related contextual document.

The recent surge in the dissemination of memes has led to an evolving body of studies on meme analysis in which the primary focus has been on tasks, such as emotion analysis (Sharma et al., 2020a), visual-semantic role labeling (Sharma et al., 2022b), and detection of sarcasm, hate-speech (Kiela et al., 2020), trolling (Hegde et al., 2021) and harmfulness (Pramanick et al., 2021; Sharma et al., 2022a).

These studies indicate that off-the-shelf multi-modal models, which perform well on several traditional visual-linguistic tasks, struggle when applied to memes (Kiela et al., 2020; Baltrušaitis et al., 2017; Sharma et al., 2022a). The primary reason behind this is the contextual dependency of memes for their accurate assimilation and analysis. Websites like `knowyourmeme.com` (KYM) facilitate important yet restricted information. MEME*X* requires the model to learn the cross-modal analogies shared by the contextual evidence and the meme at various levels of information abstraction, towards detecting the crucial explanatory evidence[1]. The critical challenge is to represent the abstraction granularity aptly. Therefore, we formulate MEME*X* as an "evidence detection" task, which can help deduce pieces of contextual evidence that help bridge the abstract gap. However, besides including image and text modality, there is a critical need to inject contextual signals that compensate for the constraints due to the visual-linguistic grounding offered by conventional approaches.

Motivated by this, we first curate `MCC`, a new dataset that captures various memes and related contextual documents. We also systematically experiment with various multimodal solutions to address MEME*X*, which culminates into a novel framework, named `MIME` (MultImodal Meme Explainer). Our model primarily addresses the challenges posed by the knowledge gap and multimodal abstraction and delivers optimal detection of contextual evidence for a given pair of memes and related contexts. In doing so, `MIME` surpasses several competitive and conventional baselines.

To summarize, we make the following main contributions[2]:

- **A novel task**, MEME*X*, aimed to identify explanatory evidence for memes from their related contexts.
- **A novel dataset**, `MCC`, containing 3400 memes and related context, along with gold-standard human annotated evidence sentence-subset.
- **A novel method**, `MIME` that uses common sense-enriched meme representation to identify evidence from the given context.
- **Empirical analysis** establishing `MIME`'s superiority over various unimodal and multimodal baselines, adapted for the MEME*X* task.

[1]A comparative analysis for KYM and `MIME` is presented in Appendix C.

[2]We provide a sample subset from the `MCC` dataset and the source code along with this manuscript.

2 Related Work

This section briefly discusses relevant studies on meme analysis that primarily attempt to capture a meme's affective aspects, such as *hostility* and *emotions*. Besides these, we also review other popular tasks to suitably position our work alongside different related research dimensions being explored.

Meme Analysis: Several shared tasks have been organized lately, a recent one on detecting heroes, villains, and victims from memes (Sharma et al., 2022b). Other similar initiatives include troll meme classification (Suryawanshi and Chakravarthi, 2021) and meme-emotion analysis via their sentiments, types and intensity prediction (Sharma et al., 2020b). Notably, hateful meme detection was introduced by Kiela et al. (2020) and later followed-up by Zhou et al. (2021). Significant interest was garnered as a result of these, wherein various models were developed. A few efforts included fine-tuning Visual BERT (Li et al., 2019a), and UNITER (Chen et al., 2020), along with using Detectron-based representations (Velioglu and Rose, 2020; Lippe et al., 2020) for hateful meme detection. On the other hand, there were systematic efforts involving unified and dual-stream encoders using Transformers (Muennighoff, 2020; Vaswani et al., 2017a), ViLBERT, VLP, UNITER (Sandulescu, 2020; Lu et al., 2019; Zhou et al., 2020; Chen et al., 2020), and LXMERT (Tan and Bansal, 2019) for dual-stream ensembling. Besides these, other tasks addressed anti-semitism (Chandra et al., 2021), propaganda techniques (Dimitrov et al., 2021), harmfulness (Pramanick et al., 2021), and harmful targets in memes (Sharma et al., 2022a) .

Visual Question Answering (VQA): Early prominent work on VQA with a framework encouraging *open-ended* questions and candidate answers was done by Antol et al. (2015). Since then, there have been multiple variations observed. Antol et al. (2015) classified the answers by jointly representing images and questions. Others followed by examining cross-modal interactions via attention types not restricted to co/soft/hard-attention mechanisms (Lu et al., 2016; Anderson et al., 2018; Malinowski et al., 2018), effectively learning the explicit correlations between question tokens and localised image regions. Notably, there was a series of attempts toward incorporating common-sense reasoning (Zellers et al., 2019; Wu et al., 2016, 2017; Marino et al., 2019). Many of these studies also leveraged information from external knowledge bases for addressing VQA tasks. General models like UpDn (Anderson et al., 2018) and LXMERT (Tan and Bansal, 2019) explicitly leverage non-linear transformations and Transformers for the VQA task, while others like LMH (Clark et al., 2019) and SSL (Zhu et al., 2021) addressed the critical language priors constraining the VQA performances, albeit with marginal enhancements.

Cross-modal Association: Due to an increased influx of multimodal data, the cross-modal association has recently received much attention. For cross-modal retrieval and vision-language pretraining, accurate measurement of cross-modal similarity is imperative. Traditional techniques primarily used concatenation of modalities, followed by self-attention towards learning cross-modal alignments (Wang et al., 2016). Following the object-centric approaches, Zeng et al. (2021) and Li et al. (2020) proposed a multi-grained alignment approach, which captures the relation between visual concepts of multiple objects while simultaneously aligning them with text and additional meta-data. On the other hand, several methods also learned alignments between coarse-grained features of images and texts while disregarding object detection in their approaches (Huang et al., 2020; Kim et al., 2021). Later approaches attempted diverse methodologies, including cross-modal semantic learning from visuals and contrastive loss formulations (Yuan et al., 2021; Jia et al., 2021; Radford et al., 2021).

Despite a comprehensive coverage of cross-modal and meme-related applications in general, there are still several fine-grained aspects of memes like *memetic contextualization* that are yet to be studied. Here, we attempt to address one such novel task, MEMEX.

3 MCC: Meme Context Corpus

Due to the scarcity of publicly-available large-scale datasets that capture memes and their contextual information, we build a new dataset, MCC (Meme Context Corpus). The overall dataset curation was conducted in three stages: (i) meme collection, (ii) content document curation, and (iii) dataset annotation. These stages are detailed in the remaining section.[3]

	Annotation Guidelines
1	Meme and the associated context should be understood before annotation.
2	Meme's semantics must steer the annotation.
3	Self-contained, minimal units of information can constitute evidence.
4	Valid evidence may or may not occur contiguously.
5	Cases not supported by a contextual document should be searched on other established sources.
6	Ambiguous cases should be skipped.

Table 2: Prescribed guidelines for MCC's annotation.

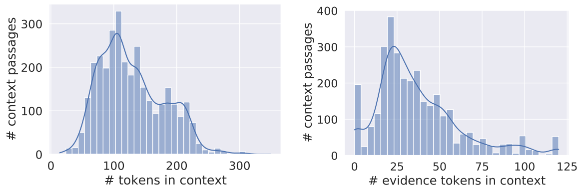
3.1 Meme Collection

In this work, we primarily focus on *political* and *historical*, *English language* memes. The reason for such a choice is the higher presence of online memes based on these topics. This is complemented by the availability of systematic and detailed information documented over well-curated digital archives. In addition to these categories, we also extend our search-space to some other themes pertaining to *movies* and *entertainment* as well. For scraping the meme images, we mainly leverage Google Images[4] and Reddit[5], for their extensive search functionality and multimedia diversity.

3.2 Context Document Curation

We curate contextual corpus corresponding to the memes collected in the first step. This context typically constitutes pieces of evidence for the meme's background context, towards which we consider Wikipedia[6] (*Wiki*) as a primary source. We use a Python-based wrapper API[7] to obtain text from Wikipedia pages. For example, for *Trump*, we crawl his Wiki. page [8]. For the scenarios wherein sufficient details are not available on a page, we look for fine-grained Wiki topics or related *non-Wiki* news articles. For several other topics, we explore community-based discussion forums and question-answering websites like Quora[9] or other general-purpose websites.

[3] Additional details are included in Appendix B.
[4] https://www.google.com/imghp
[5] https://www.reddit.com/
[6] https://www.wikipedia.org/
[7] https://github.com/goldsmith/Wikipedia
[8] https://en.wikipedia.org/wiki/Donald_Trump
[9] https://www.quora.com/



(a) Context size distribution (b) Evidence size distribution

Figure 1: Distribution of # tokens (n) in: (a) related contexts ($n \in [14, 349]$) and (b) context evidences ($n \in [5, 312]$) (outliers > 125, not depicted).

3.3 Annotation Process

Towards curating MCC, we employed *two* annotators, one male and the other female (both Indian origin), aged between 24 to 35 yrs, who were duly paid for their services, as per Indian standards. Moreover, both were professional lexicographers and social media savvy, well versed in the urban social media vernacular. A set of prescribed guidelines for the annotation task, as shown in Table 2, were shared with the annotators. Once the annotators were sure that they understood the meme's background, they were asked to identify the sentences in the context document that succinctly provided the background for the meme. We call these sentences "evidence sentences" as they facilitate (sub-)optimal *evidences* that constitute likely background information. The annotation quality was assessed using *Cohen's Kappa*, after an initial dry-run and the final annotation. The *first* stage divulged a *moderate* agreement score of 0.55, followed by several rounds of discussions, leading to a *substantial* agreement score of 0.72.

3.4 Dataset Description

The topic-wise distribution of the memes reflects their corresponding availability on the web. Consequently, MCC proportionately includes History (38.59%), Entertainment (15.44%), Joe Biden (12.17%), Barack Obama (9.29%), Coronavirus (7.80%), Donald Trump (6.61%), Hillary Clinton (6.33%), US Elections (1.78%), Elon Musk (1.05%) and Brexit (0.95%). Since the contextual document-size corresponding to the memes was significantly large (on average, each document consists of 250 sentences), we ensured tractability within the experimental setup by limiting the scope of the meme's related context to a subset of the entire document. Upon analyzing the token distribution for the ground-truth pieces of evidence, we observe the maximum token length of 312 (c.f. Fig. 1b for the evidence token distribution). Therefore, we set the maximum context length threshold to 512 tokens. This leads to the consideration of an *average* of ≈ 128 tokens and a *maximum* of over 350 tokens (spanning 2-3 paragraphs) within contextual documents (c.f. Fig. 1a for the context token distribution). This corresponds to a maximum of 10 sentences per contextual document.

We split the dataset into 80:10:10 ratio for train/validation/test sets, resulting in 3003 memes in the *train* set and 200 memes each in *validation* and *test* sets. Moreover, we ensure proportionate distributions among the train, test, and validation sets. Each sample in MCC consists of a meme image, the context document, OCR-extracted meme's text, and a set of evidence sentences.

4 Methodology

In this section, we describe our proposed model, MIME. It takes a meme (an image with overlaid text) and a related context as inputs and outputs a sequence of labels indicating whether the context's constituting *evidence sentences*, either in part or collectively, explain the given meme or not.

As depicted in Fig. 2, MIME consists of a text encoder to encode the context and a multimodal encoder to encode the meme (image and text). To address the complex abstraction requirements, we design a Knowledge-enriched Meme Encoder (KME) that augments the joint multimodal representation of the meme with external common-sense knowledge via a gating mechanism. On the other hand, we use a pre-trained BERT model to encode the sentences from the candidate context.

For designing a multi-layered contextual-enrichment pipeline, we first set up a Meme-Aware Transformer (MAT) to integrate meme-based information into the context representation. Next, we design a Meme-Aware LSTM (MA-LSTM) that sequentially processes the context representations conditioned upon the meme-based representation. Lastly, we concatenate the last hidden representations of the context from MA-LSTM and the meme representation and use this jointly-contextualized meme representation for evidence detection. Below, we describe each component of MIME in detail.

Context Representation: Given a related context C consisting of sentences $[c_1, c_2 ... c_n]$, we encode each sentence in C *individually* using a pre-trained BERT encoder, and the pooled output corresponding to the [CLS] token is used as the context

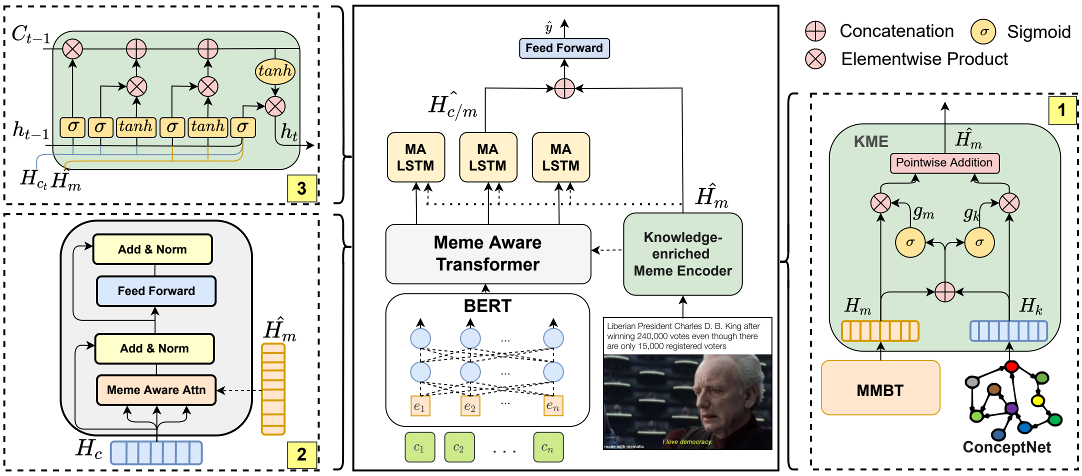


Figure 2: The architecture of our proposed model, MIME. We obtain external knowledge-enriched multimodal meme representation using Knowledge-enriched Meme Encoder (KME 1). We make use of a Meme-Aware Transformer (MAT 2) and a Meme-Aware LSTM layer (MA-LSTM 3) to incorporate the meme information while processing the context smoothly.

representation. Finally, we concatenate the individual sentence representation to get a unified context representation $H_c \in \mathbb{R}^{n \times d}$, with a total of n sentences.

Knowledge-enriched Meme Encoder: To incorporate external knowledge, we use ConceptNet (Speer et al., 2016) knowledge graph (KG) as a source of external commonsense knowledge. To take full advantage of the KG, and at the same time to avoid the query computation cost, we use the last layer from a pre-trained graph convolutional network (GCN), trained over ConceptNet (Malaviya et al., 2020).

We first encode meme M by passing the meme image M_i and the meme text M_t[10] to an empirically designated pre-trained model MMBT (Kiela et al., 2019), to obtain a multimodal representation of the meme $H_m \in \mathbb{R}^d$. Next, to get the external knowledge representation, we obtain the GCN node representation corresponding to the words in the meme text M_t. This is followed by average-pooling these embeddings to obtain the unified knowledge representation $H_k \in \mathbb{R}^d$.

To learn a knowledge-enriched meme representation $\hat{H}_m$, we design a Gated Multimodal Fusion (GMF) block. As part of this, we employ a *meme gate* (g_m) and the *knowledge gate* (g_k), to modulate and fuse the corresponding representations.

$$\begin{aligned} g_m &= \sigma([H_m + H_k]W_m + b_m) \\ g_k &= \sigma([H_m + H_k]W_k + b_k) \end{aligned} \tag{1}$$

[10]Extracted using Google Vision's OCR API: `https://cloud.google.com/vision/docs/ocr`

Here, W_m and $W_k \in \mathbb{R}^{2d \times d}$ are trainable parameters.

Meme-Aware Transformer: A conventional Transformer encoder (Vaswani et al., 2017b) uses self-attention, which facilitates learning of the inter-token contextual semantics. However, it does not consider any additional contextual information that might be useful for generating the query, key, and value representations. Inspired by the context-aware self-attention proposed by Yang et al. (2019), in which the authors proposed several ways to incorporate *global*, *deep* and *deep-global* contexts while computing self-attention over *embedded textual tokens*, we propose a meme-aware multi-headed attention (MHA). This facilitates the integration of *multimodal meme information* while computing the self-attention over context representations. We call the resulting encoder a meme-aware Transformer (MAT) encoder, which is aimed at computing the cross-modal affinity for H_c, conditioned upon the knowledge-enriched meme representation $\hat{H}_m$.

Conventional self-attention uses query, key, and value vectors from the same modality. In contrast, as part of meme-aware MHA, we first generate the key and the value vectors conditioned upon the meme information and then use these vectors via conventional multi-headed attention-based aggregation. We elaborate on the process below.

Given the context representation H_c, we first calculate the conventional query, key, and value vectors Q, K, $V \in \mathbb{R}^{n \times d}$, respectively as given below:

$$[QKV] = H_c[W_Q W_K W_V] \quad (2)$$

Here, n is the maximum sequence length, d is the embedding dimension, and W_Q, W_K, and $W_V \in \mathbb{R}^{d\times d}$ are learnable parameters.

We then generate new key and value vectors $\hat{K}$ and $\hat{V}$, respectively, which are conditioned on the meme representation $\hat{H}_m \in \mathbb{R}^{1\times d}$ (broadcasted corresponding to the context size). To regulate the memetic and contextual interaction, we use a gating parameter $\lambda \in \mathbb{R}^{n\times 1}$. Here, U_k and U_v constitute learnable parameters.

$$\begin{bmatrix}\hat{K}\\ \hat{V}\end{bmatrix} = (1 - \begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix})\begin{bmatrix}K\\ V\end{bmatrix} + \begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix}(\hat{H_m}\begin{bmatrix}U_k\\ U_v\end{bmatrix}) \quad (3)$$

We learn the parameters λ_k and λ_v using a sigmoid based gating mechanism instead of treating them as hyperparameters as follows:

$$\begin{bmatrix}\lambda_k\\ \lambda_v\end{bmatrix} = \sigma(\begin{bmatrix}K\\ V\end{bmatrix}\begin{bmatrix}W_{k_1}\\ W_{v_1}\end{bmatrix} + \hat{H_m}\begin{bmatrix}U_k\\ U_v\end{bmatrix}\begin{bmatrix}W_{k_2}\\ W_{v_2}\end{bmatrix}) \quad (4)$$

Here, W_{k_1}, W_{v_1}, W_{k_2} and $W_{v_2} \in \mathbb{R}^{d\times 1}$ are learnable parameters.

Finally, we use the query vector Q against $\hat{K}$ and $\hat{V}$, conditioned on the meme information in a conventional scaled dot-product-based attention. This is extrapolated via multi-headed attention to materialize the Meme-Aware Transformer (MAT) encoder, which yields meme-aware context representations $H_{c/m} \in \mathbb{R}^{n\times d}$.

Meme-Aware LSTM: Prior studies have indicated that including a recurrent neural network such as an LSTM with a Transformer encoder like BERT is advantageous. Rather than directly using a standard LSTM in `MIME`, we aim to incorporate the meme information into sequential recurrence-based learning. Towards this objective, we introduce Meme-Aware LSTM (MA-LSTM) in `MIME`. MA-LSTM is a recurrent neural network inspired by (Xu et al., 2021) that can incorporate the meme representation $\hat{H}_m$ while computing cells and hidden states. The gating mechanism in MA-LSTM allows it to assess how much information it needs to consider from the hidden states of the enriched context and meme representations, $H_{c/m}$ and $\hat{H}_m$, respectively.

Fig. 2 shows the architecture of MA-LSTM. We elaborate on the working of the MA-LSTM cell below. It takes as input the previous cell states c_{t-1}, previous hidden representation h_{t-1}, current cell input H_{c_t}, and an additional meme representation $\hat{H}_m$. Besides the conventional steps involved for the computation of *input, forget, output* and *gate* values w.r.t the input H_{c_t}, the *input* and the *gate* values are also computed w.r.t the additional input $\hat{H}_m$. The final *cell* state and the *hidden* state outputs are obtained as follows:

$$\begin{aligned} c_t &= f_t \odot c_{t-1} + i_t \odot \hat{c}_t + p_t \odot \hat{s}_t \\ h_t &= o_t \odot tanh(c_t) \end{aligned}$$

The hidden states from each time-step are then concatenated to produce the unified context representation $\hat{H_{c/m}} \in \mathbb{R}^{n\times d}$.

Prediction and Training Objective: Finally, we concatenate $\hat{H}_m$ and $\hat{H_{c/m}}$ to obtain a joint context-meme representation, which we then pass through a feed-forward layer to obtain the final classification. We use cross-entropy loss to optimize our model.

5 Baseline Models

We experiment with various unimodal and multimodal encoders for systematically encoding memes and context representations to establish comparative baselines. The details are presented below.

Unimodal Baselines: • **BERT (Devlin et al., 2019):** To obtain text-based unimodal meme representation. • **ViT (Dosovitskiy et al., 2020):** Pre–trained on ImageNet to obtain image-based unimodal meme representation.

Multimodal Baselines: • **Early-fusion:** To obtain a concatenated multimodal meme representation, using BERT and ViT model. • **MMBT (Kiela et al., 2019):** For leveraging projections of pre-trained image features to text tokens to encode via multimodal bi-transformer. • **CLIP (Radford et al., 2021):** To obtain multimodal representations from meme using CLIP image and text encoders, whereas CLIP text encoder for context representation. • **BAN (Kim et al., 2018):** To obtain a joint representation using low-rank bilinear pooling while leveraging the dependencies among two groups of input channels. • **VisualBERT (Li et al., 2019b):** To obtain multimodal pooled representations for memes, using a Transformer-based visual-linguistic model.

6 Experimental Results

This section presents the results (averaged over five independent runs) on our thematically diversified *test-set* and perform a comparison, followed by qualitative and error analysis. For comparison,

Type	Model	Acc.	F1	Prec.	Rec.	E-M
UM	Bert	0.638	0.764	0.768	0.798	0.485
	ViT	0.587	0.698	0.711	0.720	0.450
MM	E-F	0.646	0.772	0.787	0.798	0.495
	CLIP	0.592	0.709	0.732	0.747	0.460
	BAN	0.638	0.752	0.767	0.772	0.475
	V-BERT	0.641	0.765	0.773	0.783	0.490
	MMBT †	0.650	0.772	0.790	0.805	0.505
	MIME *	**0.703**	**0.812**	**0.833**	**0.828**	**0.585**
$\Delta_{(\text{MIME -MMBT})}$		↑ 5.34%	↑ 3.97%	↑ 4.26%	↑ 2.31%	↑ 8.00%

Table 3: Comparison of different approaches on MCC. The last row shows the absolute improvement of MIME over MMBT (the best baseline). E-F: Early Fusion and V-BERT: VisualBERT.

MMBT

John Paul Jones was a Scottish-American naval captain who was the United States' first well-known naval commander in the American Revolutionary War. He made many friends among U.S political elites, as well as enemies (who accused him of piracy). **His actions in British waters during the Revolution earned him an international reputation which persists to this day.**

MIME

John Paul Jones was a Scottish-American naval captain who was the United States' first well-known naval commander in the American Revolutionary War. He made many friends among U.S political elites, as well as enemies (who accused him of piracy). **His actions in British waters during the Revolution earned him an international reputation which persists to this day.**

Table 4: Evidence detection from MMBT (top) and MIME (bottom) for a sample meme. The emboldened sentences in blue indicate **ground-truth evidences** and the highlighted sentences indicate model prediction.

we use the following standard metrics – accuracy (Acc.), macro averaged F1, precision (Prec.), recall (Rec.), and exact match (E-M) score[11]. Additionally, as observed in (Beskow et al., 2020), we perform some basic image-editing operations like adjusting *contrast, tint, temperature, shadowing* and *highlight*, on meme images in MCC for: (i) optimal OCR extraction of meme text, and (ii) noise-resistant feature learning from images[12].

Meme-evidence Detection (MEMEX): As part of performance analysis, we observe from Table 3 that unimodal systems, in general, perform with mediocrity, with the Bert-based model yielding a relatively better F1 score of 0.7641, as compared to the worst score of 0.6985 by ViT-based model. It can be reasoned that textual cues would be significantly pivotal in modeling association when the target modality is also text-based. On the contrary, purely image-based conditioning would not be sufficient for deriving fine-grained correlations required for accurately detecting correct evidence. Also, the lower precision, as against the higher recall scores, suggests the inherent noise being additionally modelled.

On the other hand, multimodal models either strongly compete or outperform unimodal ones, with CLIP being an exception. With an impressive F1 score of 0.7725, MMBT fares optimally compared to the other comparative multimodal baselines. This is followed by the early-fusion-based approach and VisualBERT, with 0.7721 and 0.7658 F1 scores, respectively. BAN (Bilinear Attention Network) performs better than early-fusion and CLIP but falls short by a 1-2% F1 score. Models like MMBT and VisualBERT leverage pre-trained unimodal encoders like BERT and ResNet and project a systematic joint-modeling scheme for multiple modalities. Although this has proven to be beneficial towards addressing tasks that leverage visual-linguistic grounding, especially when pre-trained using large-scaled datasets like MSCOCO (VisualBERT), their limitations can be ascertained from Table 3, wherein MIME yields absolute improvements of 5.34%, 3.97%, 4.26%, 2.31% and 8.00% in accuracy, F1 score, precision, recall, and exact match scores, respectively, over the best baseline, MMBT. This suggests potential improvement that a systematic and optimal contextualization-based approach like MIME can offer.

Analysing Detected Evidences: We analyze the detected evidence by contrasting MIME's prediction quality with that of MMBT. The meme depicted in Table 4 does not explicitly convey much information and only mentions two entities, "John Paul Jones" and "The British Isles". The MMBT baseline predicts the first sentence as an explanation, which contains the word "John Paul Jones", whereas MIME correctly predicts the last sentence that explains the meme. It is interesting to observe the plausible multimodal analogy that might have led MIME to detect the relevant evidence in this case correctly. In general, we observe that the evidence predicted by MMBT does not fully explain the meme, whereas those predicted by MIME are often more fitting.

Ablation Study: MIME's key modules are Knowledge-enriched Meme Encoder (KME), Meme-Aware Transformer (MAT) encoder, and Meme-Aware LSTM (MA-LSTM). The incremental assessment of these components, over MMBT

[11]Additional experimental details are available in Appendix A.

[12]See Section 8 for details on *Terms and Conditions for Data Usage*.

System	Model	Acc.	F1	Prec.	Rec.	E-M
MMBT & variants	MMBT	0.650	0.772	0.790	0.805	0.505
	+ KME	0.679	0.789	0.804	0.822	0.550
	+ MAT	0.672	0.793	0.810	0.814	0.540
	+ MA-L	0.639	0.780	0.791	0.808	0.490
MIME & variants	– MA-L	0.694	0.800	0.826	0.8234	0.560
	– MA-L + BiL	0.689	0.807	0.8141	0.826	0.565
	– MAT	0.649	0.783	0.788	0.811	0.510
	– MAT + T	0.687	0.779	0.801	0.813	0.560
	MIME	**0.703**	**0.812**	**0.833**	**0.828**	**0.585**

Table 5: Component-wise evaluation: each component contributes to the performance of MIME, while removing them inhibits it. T: Transformer, L: LSTM, BiL: Bi-LSTM and MA: Meme-Aware.

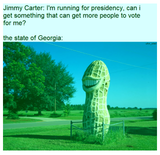
Meme	Related Context
	Heart of Darkness (1899) is a novella by Polish-English novelist Joseph Conrad. It tells the story of Charles Marlow, a sailor who takes on an assignment from a Belgian trading company as a ferry-boat captain in the African interior. **The novel is widely regarded as a critique of European colonial rule in Africa, whilst also examining the themes of power dynamics and morality. Although Conrad does not name the river where the narrative takes place, at the time of writing the Congo Free State, the location of the large and economically important Congo River, was a private colony of Belgium's King Leopold II.**
	The Jimmy Carter Peanut Statue is a monument located in Plains, Georgia, United States. **Built in 1976, the roadside attraction depicts a large peanut with a toothy grin, and was built to support Jimmy Carter during the 1976 United States presidential election.** The statue was commissioned by the Indiana Democratic Party during the 1976 United States presidential election as a form of support for Democratic candidate Jimmy Carter's campaign through that state. The statue, a 13-foot (4.0 m) peanut, references Carter's previous career as a peanut farmer.
	On February 26, 1815, Napoleon managed to sneak past his guards and somehow escape from Elba, slip past interception by a British ship, and return to France. Immediately, people and troops began to rally to the returned Emperor. French police forces were sent to arrest him, but upon arriving in his presence, they kneeled before him. Triumphantly, Napoleon returned to Paris on March 20, 1815. **Paris welcomed him with celebration, and Louis XVIII, the new king, fled to Belgium. With Louis only just gone, Napoleon moved back into the Tuileries.** The period known as the Hundred Days had begun.

Table 6: Prediction errors from MIME on three *test-set* samples. The emboldened sentences in blue indicate **ground-truth evidences** and the highlighted sentences indicate model prediction.

as a base model, can be observed from Table 5. Adding external knowledge-based cues along with the MMBT representation via KME leads to an enhancement of 0.98%-2.91% and 5% across the first four and the exact match, respectively. Similar enhancements are observed with MAT and MA-LSTM, with increments of 0.91-2.25% and 0.06-2.25%, respectively. Therefore, it can be reasonably inferred that KME, MAT, and MA-LSTM distinctly contribute towards establishing the efficacy of MIME.

On removing MA-LSTM, we notice a distinct performance drop $\in [0.47, 2.50]\%$ across all five metrics. Dropping MAT from MIME downgrades the performance by 1.67-5.38% for the first four metrics and by 7.5% for the exact match score.

Finally, we examine the influence via replacement by employing a standard Transformer-based encoder instead of MAT and a BiLSTM layer instead of MA-LSTM, in MIME. The former results in a drop of 1.45-3.28% across all five metrics. In contrast, the drop for the latter is observed to be 0.21%-2.00%. This suggests the utility of systematic memetic contextualization while addressing MEME*X*.

Error Analysis: Here, we analyze different types of errors incurred by the model. As observed from the first example in Table 6, ground-truth evidences contain abstract concepts like *power dynamics and morality*, along with various novel facts, which induce non-triviality. On the contrary, the second example depicts a partial prediction, wherein the extra excerpt detected by the MIME is likely due to the concepts of *presidential race, Jimmy Carter and visual description of the peanut statue*, possibly while leveraging its inductive bias. Finally, the model just mapped its prediction based on the embedded meme text, e.g., #3, while partly oblivious to the meme's visuals. Overall, MIME obtains an exact match for 58.50% of the test-set cases. At the same time, it cannot predict any explanation for 12.5% cases. The model obtains partial matches for about 14% of the cases, and for the remaining 14%, the model makes wrong predictions.[13]

7 Conclusion

This work proposed a new task – MEME*X* that aims to identify evidence from a given context to explain the meme. To support this task, we also curated MCC, a novel manually-annotated multimodal dataset encompassing a broad range of topics. After that, we benchmarked MCC on several competitive systems and proposed MIME, a novel modeling framework that utilizes knowledge-enriched meme representation and integrates it with context via a unique multi-layered fusion mechanism. The empirical examination and an extensive ablation study suggested the efficacy of MIME and its constituents. We then analyzed MIME's correct contextual mapping heuristics, juxtaposed with its limitations, suggesting the possible scope of improvement.

[13]Further discussion is available in Appendix D.

8 Ethics and Broader Impact

Reproducibility. We present detailed hyper-parameter configurations in Appendix A and Table 7. The source code and a sample from the MCC dataset are shared as supplementary materials. We commit to sharing the complete dataset upon the acceptance of this paper.

Data Collection. The data collection protocol was duly approved by an ethics review board.

User Privacy. The information depicted/used does not include any personal information.

Terms and Conditions for data usage. We performed basic image editing (c.f. Section 6) on the meme images downloaded from the Internet and used for our research. This ensures non-usage of the artwork/content in its original form. Also, we will release links to the memes instead of the actual memes for research purposes. In this way, we ensure that if a user deletes a posted meme, it will no longer be available in our datasets.

Moreover, we already included details of the subreddits and keywords used to collect meme content and the sources used for obtaining contextual document information as part of Appendix B.1, Section 3.2 and Figure 4d. Since the our dataset (MCC) contains material collected from various web-based sources in the public domain, the copyright and privacy guidelines applied are as specified by these corresponding sources, a few of them as follows:

- Wikipedia: Text of Creative Commons Attribution-ShareAlike 3.0.[14]
- Quora: License and Permission to Use Your Content, Section 3(c).[15]
- Reddit Privacy Policy: Personal information usage and protection.[16]
- Reddit Content Policy.[17]

Future adaptations or continuation of this work would be required to adhere to the policies prescribed herein. Finally, we will also release a data curation script for curating a new dataset in the format required by MEMEX so that it is not source dependent.

[14] https://en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License

[15] https://www.quora.com/about/tos

[16] https://www.reddit.com/policies/privacy-policy

[17] https://www.redditinc.com/policies/content-policy

Annotation. The annotation was conducted by NLP experts or linguists in India, who were fairly treated and duly compensated. We conducted several discussion sessions to ensure that all annotators understood the annotation requirements for MEMEX.

Biases. Any biases found in the dataset are unintentional, and we do not intend to cause harm to any group or individual. We acknowledge that memes can be subjective, and thus it is inevitable that there would be biases in our gold-labeled data or the label distribution. This is addressed by working on a dataset created using a diverse set of topics and following a well-defined annotation scheme, which sets explicitly characterizes meme-evidence association.

Misuse Potential. Our dataset can be potentially used for ill-intended purposes, such as biased targeting of individuals/communities/organizations, etc., that may or may not be related to demographics and other information within the text. Intervention with human moderation would be required to ensure this does not occur.

Intended Use. We curated MCC solely for research purposes, in-line with the associated usage policies prescribed by various sources/platforms. This applies in its entirety to its further usage as well. We commit to releasing our dataset, aiming to encourage research in studying complex multi-modal associations. We will distribute the dataset for research purposes only, without a license for commercial use. We believe that it represents a valuable resource when used appropriately.

Environmental Impact. Finally, large-scale models require a lot of computations, which contribute to global warming (Strubell et al., 2019). However, in our case, we do not train such models from scratch; instead, we fine-tune them on a relatively small dataset.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal machine learning: A survey and taxonomy.

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.

Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. "subverting the jewtocracy": Online antisemitism detection using multimodal deep learning.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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)*, pages 4069–4082, Hong Kong, China. 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. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

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. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Siddhanth U Hegde et al. 2021. Do images really do the talking? Analysing the significance of images in Tamil troll meme classification. *arXiv:2108.03886*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text.

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. *Advances in Neural Information Processing Systems*, 33.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, pages 13–23. Curran Associates, Inc.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 289–297, Red Hook, NY, USA. Curran Associates Inc.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Choi Yejin. 2020. Commonsense knowledge base completion with structural and semantic context. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:2925–2933.

Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. 2018. Learning visual question answering by bootstrapping hard attention.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes.

Shraman Pramanick et al. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *EMNLP (Findings)*, pages 4439–4455.

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.

Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020a. Semeval-2020 task 8: Memotion analysis – the visuo-lingual metaphor!

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020b. 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.

Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2022a. DISARM: Detecting the victims targeted by harmful memes. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1572–1588, Seattle, United States. Association for Computational Linguistics.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3645–3650, Florence, Italy.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.

Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1367–1381.

Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4622–4630.

Lu Xu, Zhanming Jie, Wei Lu, and Lidong Bing. 2021. Better feature integration for named entity recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3457–3469, Online. Association for Computational Linguistics.

Baosong Yang, Jian Li, Derek Wong, Lidia Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:387–394.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.

Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2021. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Modality	Model	BS	EP	# Param (M)	Runtime (s)
UM	Bert	16	20	110	0.66
	ViT			86	0.64
MM	Early Fusion			196	0.62
	CLIP			152	0.73
	BAN			200	0.75
	VisualBERT			247	0.78
	MMBT			279	0.99
	MIME			303	0.72

Table 7: Hyper-parameters and per-batch *inference* runtime for each model.

A Implementation Details and Hyperparameter values

We train all the models using Pytorch on an NVIDIA Tesla V100 GPU with 32 GB dedicated memory, CUDA-11.2 and cuDNN-8.1.1 installed. For the unimodal models, we import all the pre-trained weights from the TORCHVISION.MODELS subpackage of the PyTorch framework. We initialize the remaining weights randomly using a zero-mean Gaussian distribution with a standard deviation of 0.02.

We primarily perform manual fine-tuning, over five independent runs, towards establishing an optimal configuration of the hyper-parameters involved. Finally, we train all models we experiment with using the Adam optimizer and a binary cross entropy loss as the objective function.

B Additional details about MCC

B.1 Meme Collection

For every category, we use carefully constructed search queries to obtain relevant memes from the Google Images search engine[18]. Towards searching variants for the topics related *Joe Biden*, some search queries used were “Joe Biden Political Memes”, “Joe Biden Sexual Allegation Memes”, “Joe Biden Gaffe Memes”, “Joe Biden Ukraine Memes”, among others; for memes related to *Hillary Clinton*, we had “Hillary Clinton Email Memes”, “Hillary Clinton Bill Clinton Memes”, “Hillary Clinton US Election Memes”, “Hillary Clinton President Memes”, etc. For crawling and downloading these images, we use Selenium[19], a Python framework for web browser automation.

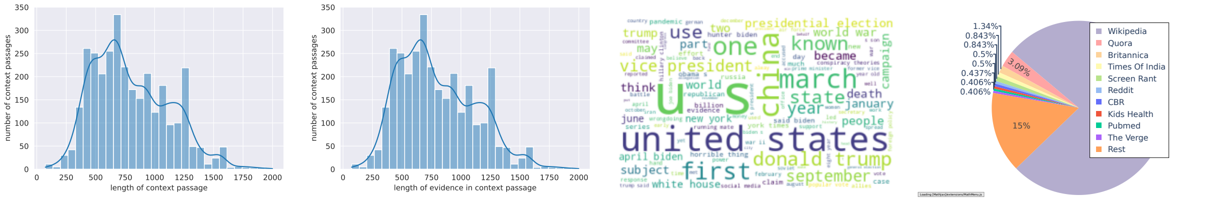
Additionally, for certain categories, we also crawl memes off Reddit. Specifically, We focus on `r/CoronavirusMemes`, `r/PoliticalHumor`, `r/PresidentialRace` subreddits. Instead of using the Python Reddit API Wrapper (PRAW), we use the Pushshift API[20], which has no limit on the number of memes crawled. For coronavirus, we crawl all memes from 1st November 2019 to 9th March 2021. For Biden, Trump, etc., we crawl memes from the other two subreddits and use a set of search queries, a subset of the overall queries we utilized. After scraping all possible memes, we perform de-duplication using dupeGuru[21], a cross-platform GUI tool to find duplicate files in a specified directory. This eliminates intra- and inter-category overlaps. We then remove any meme which is either unimodal, i.e., memes having only images (c.f. Fig. 3 (c)), or text-only blocks (c.f. Fig. 3 (a)). Additionally, to ensure further tractability of our setup, we manually filter out code-mixed (c.f. Fig. 3 (b)) and code-switched memes and memes in languages other than English. Annotating multilingual memes can be a natural extension of our work. We further segregate memes which have cartoons/animations (c.f. Fig. 3 (d)). We also filter out memes that are poor in image quality, have low resolution, etc.

Figure 3: Examples of discarded meme types: (a) Text-only, (b) Code-mixed, (c) Image-only and (d) Cartoon.

B.2 Annotation Process

Two annotators annotated the dataset. One of the annotators was male, while the other was female, and their ages ranged from 24 to 35. Moreover,

[18] https://images.google.com/
[19] https://github.com/SeleniumHQ/selenium
[20] https://github.com/pushshift/api
[21] https://github.com/arsenetar/dupeguru



(a) Context length dist. (b) Evidence length dist. (c) Word Cloud (d) Context source dist.

Figure 4: Distribution of attributes in MCC. The total number of sentences in context passage range between 2 and 16 and the number of evidence sentences in context range between 1 and 10. The most common source of context is Wikipedia

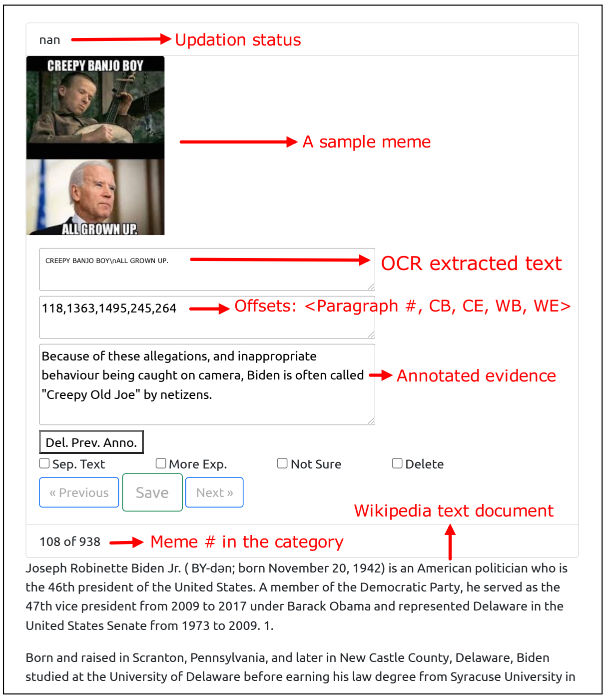


Figure 5: A Screenshot of the Annotation Tool. Abbr. details for various offsets captured: CB: Character beginning, CE: Character end, WB: Word beginning, WE: Word end.

both of them were professional lexicographers and social media savvy. Before starting the annotation process, they were briefed on the task using detailed guidelines.

For performing annotations, we build an annotation platform using JQuery[22] and Flask[23]. A screenshot of the platform is given in Fig. 5. The status of the annotation is displayed at the top. It shows a "nan" for now since the image has not been saved yet; after saving, the status is updated to "updated". Below the status, the meme is displayed. There are three text boxes: the first text box (write-able) is for the OCR text (the annotators can correct and edit the text returned by the OCR pipeline). The other two text boxes are for the offsets and the selected text. The text document in which the explanations are present is at the bottom of the page. When selecting a relevant excerpt from the document, the offsets and selected text are automatically captured and supplemented to the text boxes mentioned above. The format of the offsets, as depicted in Fig. 5 is `<Paragraph Number, Beginning Character Offset, Ending Character Offset, First Word Offset, Last Word Offset>`.

B.3 Analysis and description of MCC

It can be observed from Fig. 4d that the highest proportion of Wikipedia-based sources, followed by smaller proportions for the alternatives explored like Quora, Britannica, Times of India, etc. Additionally, the word cloud depicted in Fig. 4c suggests that most memes are about *prominent US politicians, history, and elections*. Also, context length distribution, as depicted in Fig. 4a, suggests an *almost* randomly distributed context length (in chars), with very few contexts having lengths lesser than ≈ 100 and more than ≈ 800 chars. Whereas, Fig. 4b depicts evidence length distribution, according to which a majority of pieces of evidence contain fewer than 400 characters. This corroborates the brevity of the annotated pieces of evidence from diverse contexts.

B.4 Thematic Analysis From Meme Text

We perform thematic analysis of the memetic content, using just the text embedded within memes. We took the OCR extracted meme's text and project top-20 topics using *BERTopic* (Grootendorst, 2022), which is a neural topic modeling approach with a class-based TF-IDF procedure.

We depict 0-based topic indexes and thematic keywords as 0–History, 1–Covid-19, 2–Politics, 3–

[22] https://github.com/jquery/jquery
[23] https://github.com/pallets/flask

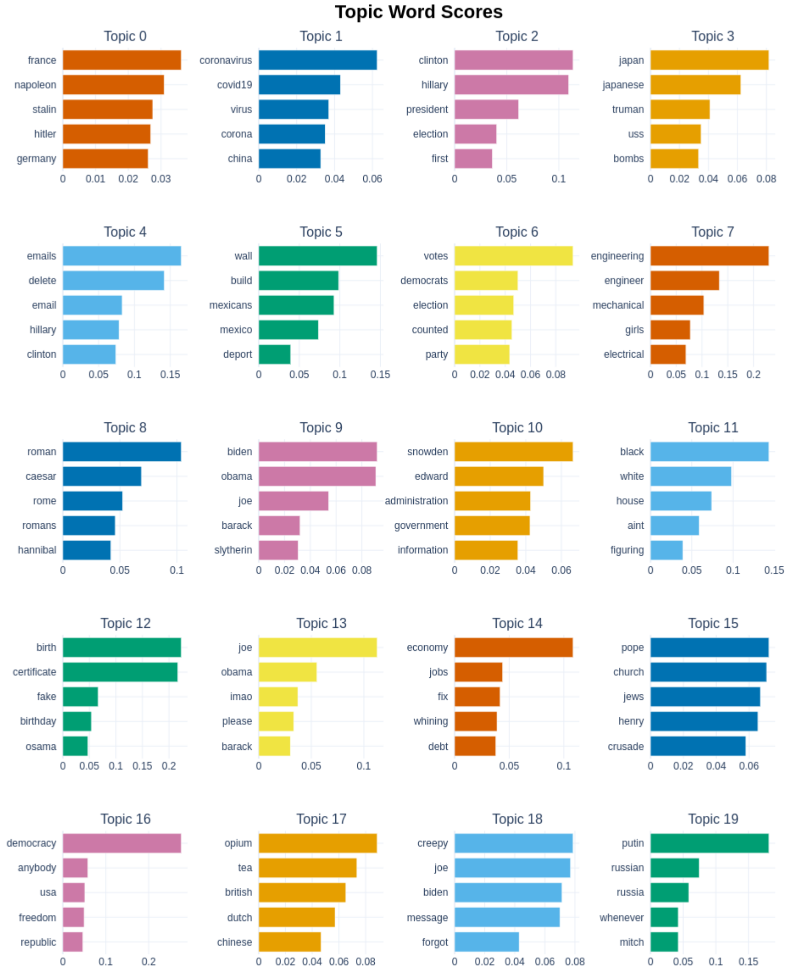


Figure 6: Top-20 prominent topics representing themes of the memetic content in MCC

KYM (knowyourmeme.com)

SpongeBob SquarePants - Steppin On the Beach (Image Details)

- 273 views
- Uploaded 3 years ago
- Origin Entry: SpongeBob SquarePants (☞ SpongeBob SquarePants is a long-running American television series created by Stephen Hillenburg, airing on Nickelodeon. Learn more on KYM.)
- Source: Reddit
- Tags: spongebob, squarepants, dday, history, normandy, landings, neverforget, steppin on the beach, world war ii, france, today in history, nickelodeon, cartoon, reddit, dankmemes
- About the Uploader: [redacted], Sr. Researcher & Scrapbooker & Media Chauffeur

MIME

The Normandy landings were the landing operations and associated airborne operations on Tuesday, 6 June 1944 of the Allied invasion of Normandy in Operation Overlord during World War II. Often referred to as D-Day, it was the largest seaborne invasion in history. The operation began the liberation of France and laid the foundations of the Allied victory on the Western Front.

Table 8: Comparison of the contextual insights obtained from KYM (knowyourmeme.com, *top*) and the one generated by MIME (*bottom*) for a sample meme. Text blacked-out ([redacted]) is for obscuring the user's identity; Emboldened sentences in blue indicate **ground-truth evidences** and the highlighted sentences indicate model prediction.

War with Japan, etc., in Fig. 6. These topics are collectively referenced and described via the most likely keywords appearing for that particular topic. This depiction also highlights how generalizable our proposed approach is in optimally detecting accurate evidence from within a given related context on various topics. Besides different high-level topics, MCC also captures the diversity in terms of the sub-topics. Although, except for a few topics like Topics: 15 and 18, reasonably diverse memes can be found in MCC.

C Comparing contexts from KYM and MIME

Here, we compare the insights available on knowyourmeme.com (also referred to by KYM) and the ones generated by our proposed modeling framework MIME, about a particular meme. For comparison, we consider a sample meme (c.f. Table 8) from our test set, which also happens to be available on KYM[24]. This meme is about a soldier (portrayed via character *SpongeBob*) stepping onto the beach on June 6th, 1944, which is an implicit reference to the D-Day landings during World War II. We present our comparative analysis in the following subsections.

C.1 MIME

Since Wikipedia articles are supposed to document in-depth factual details related to events, people, places, etc., one can expect the information obtained to be exhaustive, which is what MIME aims to leverage. MIME achieves this by establishing a cross-modal evidence-level association between memes and a supplementary context document. While there are different levels of details (with varying relatedness) present within Wikipedia documents, there are one or more sentences that *distinctly complement* the meme's intended message.

In this case, the excerpts emboldened and highlighted contribute to building the meme's rationale, as depicted in Table 8. The key advantages to using an approach like MIME can be enlisted as follows:

- Information derived can facilitate comprehensive assimilation of the meme's intended message.
- MIME does not rely on manually archived details and meta-data. Instead, it presumes the availability of a *related* context, which can be easily mined from the web.
- Finally, MIME can optimally detect accurate contextual evidence about a meme without presenting information that might not be useful.

Although MIME in its current stage has limitations, it would require active fine-tuning and optimization towards regulating its cross-modal associativity, which required memetic contextualization.

C.2 KYM

On the other hand, as can be observed from Table 8, KYM divulges the details like (a) total views, (b) time of upload, (c) origin details, (d) source, (e) relevant tags and (e) up-loader details. Most of this information could be considered as meta-data, w.r.t. the meme (template). Such information captures the details related to its multimedia-based aspects. The following factors characterize such information:

- The *origin* information about a meme is likely to be one of the most critical information, as it presents details regarding the inception of a particular meme, which is often imperative to establish the underlying perspective conveyed within a meme.
- Although *tags* aggregates a comprehensive set of related digital artifacts, it can also include some irrelevant objects.
- Other available meta-data like *no. of views, date uploaded*, etc., could be beneficial w.r.t. detecting meme's harmfulness or virality over social media, but not as much towards divulging meme's intended message.

Information provided by KYM *may or may not* be sufficient to comprehend the actual meme's intended message, as it significantly relies on human intervention towards curating such data and is always bound to be limited. Still, information like the *origin details* and *related tags* can facilitate establishing the mappings across layers of abstraction that memes typically require.

D Limitations

Although our approach, MIME is empirically observed to outperform several other competitive baselines, we do observe some limitations in the modeling capacity towards MemeX. As depicted in Table 6, there are three possible scenarios of ineffective detection – (a) no predictions, (b) partial match, and (c) incorrect predictions. The key challenges stem from the limitations in modeling the complex level of abstractions that a meme exhibits. These are primarily encountered in either of the following potential scenarios:

- A critical, yet a cryptic piece of information within memes, comes from the visuals, which typically requires some systematic integration of factual knowledge, that currently lacks in MIME.
- Insufficient textual cues pose challenges for MIME, for learning the required contextual associativity.
- Potentially spurious pieces of evidence being picked up due to the lexical biasing within the related context.

[24] https://knowyourmeme.com/photos/1500530-spongebob-squarepants