Report for Analyzing hateful Memes

Pratyush Jena

2022111016

1 Summary

This paper explores the role of memes in social media, emphasizing their potential to convey complex ideas through visual and linguistic elements. The authors introduce a novel task called MEMEX, aiming to automatically derive contextual evidence from related content to enhance meme understanding. They propose a dataset (MCC) and a framework (MIME) to address the contextual knowledge gap and multimodal challenges in meme analysis. The goal is to detect explanatory evidence for memes and related contexts, with potential applications in content moderation for governments and organizations. The authors present empirical analyses demonstrating MIME's superiority over various baselines in the MEMEX task. Overall, the contributions include a new task, dataset, and framework to improve meme comprehension in social media.

2 MEME context corpus

2.1 Data collection

The paper describes the creation of a new dataset called MCC (Meme Context Corpus) due to the limited availability of large-scale datasets capturing memes and their contextual information. The dataset creation involves three stages: meme collection, context document curation, and dataset annotation. The focus is on political and historical English-language memes, with additional themes from movies, geopolitics, and entertainment. Contextual corpus corresponding to memes is curated using sources like Wikipedia, community-based forums, and question-answering websites. The annotation process involves two professional lexicographers identifying sentences in the context document that provide background information for the memes. The dataset is topic-wise distributed, with the majority covering history, entertainment, and specific political figures. The annotation quality is assessed using Cohen's Kappa, resulting in a substantial agreement score. The dataset is split into train, validation, and test sets, with each sample containing a meme image, context document, OCR-extracted meme text, and ground truth evidence sentences. The dataset aims to facilitate research in meme understanding and analysis.

2.2 Methodology

2.2.1 Meme Collection

The paper introduces the creation of a new dataset named MCC (Meme Context Corpus) due to the limited availability of large-scale datasets capturing memes and their contextual information. The dataset curation involves meme collection, content document curation, and dataset annotation, primarily focusing on political, historical, and English-language memes.

2.2.2 Context Document Curation

Contextual corpus corresponding to the memes is curated using sources like Wikipedia, community-based forums, and question-answering websites. The annotation process involves two professional lexicographers identifying sentences in the context document that provide background information for the memes.

2.2.3 Annotation Process

Prescribed guidelines are provided for annotators to identify sentences in the context document that succinctly provide background information for the memes. The annotation quality is assessed using Cohen's Kappa, resulting in a substantial agreement score.

2.2.4 Dataset Description

The dataset, MCC, is distributed based on topics, with the majority covering history, entertainment, and specific political figures. It is split into train, validation, and test sets, with each sample containing a meme image, context document, OCR-extracted meme text, and ground truth evidence sentences.

3 Model Architecture - MIME

3.0.1 Overview

The proposed model, MIME (Multimodal Meme Explainer), takes a meme and a related context as inputs and outputs a sequence of labels indicating whether the context's constituting evidence sentences explain the given meme or not.

3.0.2 Architecture

1. **Context Representation** Given a related context, BERT is used to encode each sentence individually, and the pooled output is used as the context representation.
2. **Knowledge-enriched Meme Encoder** A Knowledge-enriched Meme Encoder (KME) is introduced to augment the multimodal meme representation with external common-sense knowledge using ConceptNet through a Gated Multimodal Fusion block.
3. **Meme-Aware Transformer** A Meme-Aware Transformer (MAT) is designed to integrate meme-based information into the context representation using a meme-aware multi-headed attention mechanism.
4. **Meme-Aware LSTM** A Meme-Aware LSTM (MA-LSTM) is introduced, inspired by previous studies, to incorporate meme information into sequential recurrence-based learning.
5. **Prediction and Training Objective** The model concatenates the enriched meme and context representations and uses a feed-forward layer to obtain the final classification. The cross-entropy loss is employed for optimization.

4 Baseline Models

In this section, the paper outlines various unimodal and multimodal encoders used as baseline models for encoding memes and context representations, serving as a basis for comparative analysis.

4.1 Unimodal Baselines

4.1.1 BERT

The BERT model [**?**] is employed to obtain text-based unimodal meme representations.

4.1.2 ViT

The Vision Transformer (ViT) model [**?**], pre-trained on ImageNet, is utilized to obtain image-based unimodal meme representations.

4.2 Multimodal Baselines

4.2.1 Early-fusion

An early-fusion approach is employed to obtain a concatenated multimodal meme representation, utilizing both the BERT and ViT models.

4.2.2 MMBT

The MMBT model [**?**] is utilized, leveraging projections of pre-trained image features to text tokens to encode via a multimodal bi-transformer.

4.2.3 CLIP

The CLIP model [**?**] is employed to obtain multimodal representations from memes. It uses CLIP image and text encoders, with the CLIP text encoder utilized for context representation.

4.2.4 BAN

The Bilinear Attention Network (BAN) model [**?**] is used to obtain a joint representation by leveraging low-rank bilinear pooling. This approach takes advantage of dependencies among two groups of input channels.

4.2.5 VisualBERT

The VisualBERT model [**?**] is employed to obtain multimodal pooled representations for memes. It utilizes a Transformer-based visual-linguistic model.

5 Experimental Results

In this section, the experimental results of the proposed model MIME are presented, along with a comparison to baseline models. The evaluation metrics include accuracy (Acc.), macro-averaged F1 score, precision (Prec.), recall (Rec.), and exact match (E-M) score, averaged over five independent runs.

5.1 Model Comparison

The results in Table 3 show a comparison of different models on the MCC dataset. Unimodal systems, such as Bert and ViT, exhibit mediocre performance. Multimodal models, including MMBT, Early Fusion (E-F), BAN, and VisualBERT, generally outperform unimodal models, with MMBT showing the best performance among baselines.

Table 1: Model Comparison on MCC Dataset

Type	**Model**	**Acc.**	**F1**	**Prec.**	**Rec.**	**E-M**
UM	Bert	0.638	0.764	0.768	0.798	0.485
UM	ViT	0.587	0.698	0.711	0.720	0.450
MM	E-F	0.646	0.772	0.787	0.798	0.495
MM	CLIP	0.592	0.709	0.732	0.747	0.460
MM	BAN	0.638	0.752	0.767	0.772	0.475
MM	V-BERT	0.641	0.765	0.773	0.783	0.490
MM	MMBT	0.650	0.772	0.790	0.805	0.505
MM	MIME	0.703	0.812	0.833	0.828	0.585

The table illustrates that MIME significantly outperforms all baseline models, achieving the highest scores in accuracy, F1 score, precision, recall, and exact match. The improvements over the best baseline, MMBT, range from 2.31% to 8.00% across different metrics, indicating the effectiveness of MIME in meme-evidence detection.

5.2 Evidence Detection Analysis (MEMEX)

Analyzing meme-evidence detection, MIME demonstrates superior performance compared to baselines. Unimodal models struggle with the complexity of meme contexts, while multimodal models, except for CLIP, show competitive results. MMBT performs optimally among baselines, and MIME surpasses it with substantial improvements in all metrics.

Table 2: MEMEX - Meme-Evidence Detection Analysis

Type	**Model**	**Acc.**	**F1**	**Prec.**	**Rec.**	**E-M**
UM	Bert	0.638	0.764	0.768	0.798	0.485
MM	MMBT	0.650	0.772	0.790	0.805	0.505
MM	MIME	0.703	0.812	0.833	0.828	0.585

The analysis highlights that MIME, with its systematic and optimal contextualization-based approach, achieves better accuracy, F1 score, precision, recall, and exact match scores compared to existing models.

5.3 Ablation Study

The ablation study in Table 5 systematically evaluates the contribution of each component in MIME by comparing variants with different modules added or removed. Key components include Knowledge-enriched Meme Encoder (KME), Meme-Aware Transformer (MAT), and Meme-Aware LSTM (MA-LSTM).

Table 3: Ablation Study Results

System	Model	Acc.	F1	Prec.	Rec.	E-M
MMBT	MMBT	0.650	0.772	0.790	0.805	0.505
MME	+ KME	0.679	0.789	0.804	0.822	0.550
MME	+ MAT	0.672	0.793	0.810	0.814	0.540
MME	+ MA-L	0.639	0.780	0.791	0.808	0.490
MIME	– MA-L	0.694	0.800	0.826	0.823	0.560
MIME	– MA-L + BiL	0.689	0.807	0.814	0.826	0.565
MIME	– MAT	0.649	0.783	0.788	0.811	0.510
MIME	– MAT + T	0.687	0.779	0.801	0.813	0.560
MIME		0.703	0.812	0.833	0.828	0.585

The table systematically evaluates the contribution of each component in MIME. External knowledge-based cues, attending over memes, and incorporating sequential modeling contribute significantly to MIME's performance.

5.4 Error Analysis

Error analysis in Table 6 provides insights into different types of errors incurred by MIME. Examples illustrate cases where ground-truth evidence contains abstract concepts or MIME produces partial predictions. The model also demonstrates occasional errors, such as mapping predictions based solely on embedded text, highlighting areas for potential improvement.

Table 4: Error Analysis

5.5 Discussion

The study concludes by discussing MIME's efficacy over other variants when the constituting components are considered both incrementally and decrementally. Notably, adding external common sense knowledge-based signals, attending over memes while processing context evidence sentences, and incorporating sequential modeling contribute significantly to MIME's performance.

6 Conclusion

This work introduces a novel task, MEMEX, focusing on identifying evidence from given contexts to explain memes. To support this task, the Multimodal Contextualization of Evidence (MCC) dataset is curated, offering a diverse range of topics. The study benchmarks MCC against competitive systems and introduces MIME, a novel modeling framework. MIME utilizes knowledge-enriched meme representation and integrates it with context through a unique multi-layered fusion mechanism. Empirical examination and an extensive ablation study validate the efficacy of MIME and its constituents. The analysis of correct contextual mapping heuristics and limitations provides insights into potential areas for improvement.

6.1 Limitations

While MIME outperforms several baselines, limitations in modeling capacity towards MEMEX are identified. Three possible scenarios of ineffective detection are outlined: (a) no predictions, (b) partial match, and (c) incorrect predictions. Challenges arise from the complexity of abstract information within memes, especially when it comes from visuals, insufficient textual cues, and potential spurious evidence due to lexical biasing.

6.2 Ethics and Broader Impact

6.2.1 Reproducibility

Detailed hyperparameter configurations are provided, and the source code and MCC dataset are publicly available for reproducibility.

6.2.2 Data Collection

The data collection protocol was approved by an ethics review board, ensuring compliance with ethical standards.

6.2.3 User Privacy

The information used does not include any personal information, prioritizing user privacy.

6.2.4 Terms and Conditions for Data Usage

Basic image editing is performed on meme images, ensuring non-usage of the original content. Details of subreddits, keywords, and sources are transparently disclosed. Future adaptations of this work must adhere to the specified policies and guidelines.

6.3 Annotation and Biases

6.3.1 Annotation

Conducted by NLP researchers or linguists in India, annotators were treated fairly and compensated. Extensive discussion sessions ensured a shared understanding of annotation requirements.

6.3.2 Biases

Unintentional biases may exist, given the subjective nature of memes. The dataset addresses this by incorporating a diverse set of topics and following a well-defined annotation scheme.

6.4 Misuse Potential

Acknowledging the possibility of misuse, the study notes that the ability to deduce relevant contextual, fact-oriented evidence might be exploited to modulate harmful expressions within memes. This underscores the importance of human moderation to prevent ill-intended uses of the proposed solution and dataset.

7 Weakness of the paper

1. I think their are no links to example codebases like in google colab or something like that cause while researching for the task whereever google colab link was there, I found it very Interactive thus learnt faster.
2. No links to datasets or github
3. scope of the meme is limited to political

8 Scope of improvement

1. I think scope of the memes can be increased to historical, modern sarcastic memes
2. I found the paper complex so links to study the complex terms and topics.
3. No videos link for summary
4. More graphs and points would be more interactive
5. example in google colab