Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Large Scale Narrative Analysis of Multimodal Memes

Jia Wang Peh, Ming Shan Hee, Bryan (Chen Zhengyu) Tan, Yuriel Wang Jun Long Ryan, Roy Ka-Wei Lee

Singapore University of Technology and Design (SUTD)
jiawang_peh@sutd.edu.sg, mingshan_hee@mymail.sutd.edu.sg, bryan_tan@mymail.sutd.edu.sg, yurieljunlongryan_wang@mymail.sutd.edu.sg, roy_lee@sutd.edu.sg

Abstract

Current computational approaches to meme analysis primarily focus on individual memes, with an emphasis on tasks such as hateful content detection and sentiment analysis. In contrast, corpus-level analyses, which are necessary to reveal in-depth thematic narratives embedded in meme corpora, have remained within the purview of social science research. While qualitative methods such as inductive content analysis provide deeper insights, they are labor-intensive and lack scalability. To address this gap, we introduce MemeTopicTrees (MemeTT), a zero-shot pipeline that clusters multimodal memes based on their targets, aspects, sentiments, and opinions. In addition to clustering, MemeTT generates a descriptive narrative for each cluster to provide a nuanced understanding of meme corpora. By integrating multimodal aspect-based sentiment analysis with hierarchical clustering, MemeTT automates both semantic analysis and narrative generation at scale. Evaluated on a combined pool of three datasets spanning political, public health, and defense domains, MemeTT successfully produces fine-grained clusters and coherent narratives, with narrative relevance most pronounced at the target and aspect levels. Evaluations show that the best-performing models are highly accurate at identifying meme targets and aspects, although maintaining high accuracy for opinions and sentiments remains challenging. Furthermore, while cluster distinctiveness is robust at the target level, room for improvement remains at lower clustering levels. Despite these challenges, this approach offers a scalable solution for analyzing public sentiment and discourse in meme corpora. We lay the foundation for future research on the understudied task of automated meme corpus analysis.

Code — https://github.com/Social-AI-Studio/MemeTT

Introduction

Motivation. Memes shape digital interactions, from facilitating political expression (Johann 2022) to perpetuating toxicity (Lim et al. 2024), making their analysis important for understanding public sentiment on social networks. However, existing computational approaches are mainly aimed at analyzing individual memes (Cao et al. 2023; Hee, Chong, and Lee 2023). This narrow focus neglects the overarching themes within meme corpora, which are integral to interpreting online discourse. Although valuable for content moderation, these methods fail to capture emerging thematic narratives and evolving sentiments, thereby limiting their utility for decision-makers.

Inductive content analysis is a staple of social science research for interpreting qualitative material. However, it is notoriously labor-intensive and time-consuming (Cho and Lee 2014). While this technique has been used to study limited meme corpora, the sheer scale of modern social media renders manual coding impractical. Given that more than one million Instagram posts per day referenced "meme" in 2020 (Instagram 2020), manual analysis is infeasible, hence the need for automated multimodal analysis techniques.

The analysis of meme corpora is challenging because memes are multimodal. Text and images often interdependently express ideas that neither can convey alone (Yus 2019). Interpreting this interplay requires multimodal inference (Toh et al. 2023). Similarly, effective meme clustering requires consideration of both modalities. In addition, memes often express varying stances on similar topics, complicating the distillation of coherent narratives. To the best of our knowledge, no end-to-end automated approach for large-scale, narrative-level analysis of memes currently exists.

Research Objectives. To address the challenges of large-scale multimodal meme analysis, we propose MemeTopicTrees (MemeTT), a scalable pipeline for corpus-level content analysis and narrative generation. Unlike methods that classify individual memes into predefined categories (e.g., hateful or non-hateful), MemeTT integrates Multimodal Aspect-Based Sentiment Analysis (MABSA) with hierarchical clustering to distill corpus-level narratives.

Specifically, MemeTT entails the inference of semantic quadruples following the *target-aspect-opinion-sentiment* formulation introduced by Li et al. (2023). While this quadruple structure was originally designed for text dialogues, we adapt it to multimodal memes. Additionally, we cluster these quadruples by shared elements into a hierarchy that surfaces themes at various levels of granularity. For example, COVID-19 memes can address various targets, from mask-wearing to remote work meetings. Mask-related memes can focus on aspects such as public health or aesthetics. Within the public health cluster, sentiments and opinions vary. Some memes positively portray masks as significant and effective, while others decry them as illogical.

Compared with traditional inductive content analysis, which iteratively distills meme corpora into broad content categories, MemeTT offers greater nuance and structure in both methodology and output: Clusters are hierarchically organized by targets, aspects, sentiments, and opinions, and are accompanied by informative narratives. This multi-layered structure captures both high-level patterns and fine-grained distinctions and yields descriptive opinion-level summaries. Our experiment applying MemeTT to a pool of datasets, including Harm-C (Pramanick et al. 2021a), revealed layered themes surrounding pandemic sentiment, including criticism of work-from-home mandates. As an example, a narrative generated by MemeTT is as follows:

> *The memes view universal work-from-home mandates with a negative sentiment because their practicality across professions is seen as absurdly limited.*

By offering a scalable approach to meme corpus analysis, MemeTT provides policymakers, government agencies, campaign strategists, and other public-facing organizations with actionable insights into public sentiment. When applied to the Harm-C corpus, for example, MemeTT surfaces not only prominent topics such as working from home, mask mandates, and vaccine mandates, but also the salient concerns associated with each. Vaccine-related memes include narratives questioning vaccine safety as well as narratives opposing mandates as infringements on personal freedom, which can support more targeted communication strategies. Fundamentally, MemeTT goes beyond topic-level analysis by showing not only *what* the public discusses, but also *which aspects* they endorse or criticize, and *why*.

Contributions. Our work makes three key contributions:

1. We propose MemeTT, an automated pipeline that advances beyond traditional inductive content analysis by delivering hierarchical insights. By leveraging Vision-Language Models (VLMs) and Large Language Models (LLMs), MemeTT enables scalable analysis of meme corpora, providing deep insights into public sentiment.
2. We extend MABSA to multimodal meme corpora, using VLMs for zero-shot inference of target-aspect-opinion-sentiment quadruples. This approach offers richer semantic representations compared to typical MABSA tasks that focus on aspect-sentiment pairs.
3. We validate MemeTT on a pool of three real-world meme datasets. Through human evaluations, we demonstrate its ability to represent meme corpora as hierarchical topic trees and to generate meaningful narratives.

Related Works

Inductive Content Analysis

Inductive content analysis is a qualitative method in which labels emerge organically during coding and are refined through iterative analysis and comparison of documents (Vears and Gillam 2022). Traditionally applied to texts, it has been extended to memes, including COVID-19 memes (Al Rousan, Al Harahsheh, and Al Rousan 2023). While effective for distilling themes, it is labor-intensive.

Advances in generative artificial intelligence have enabled tools like ChatGPT to support qualitative analysis, as demonstrated by its application to the analysis of forum posts on reducing sugar consumption (Bijker et al. 2024). Motivated by this, we use VLMs and LLMs to analyze memes, improving scalability and interpretive richness beyond traditional methods.

Aspect-Based Sentiment Analysis (ABSA)

Aspect-Based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis technique that determines sentiments toward specific entities or their characteristics, referred to as aspects (Hua et al. 2024). Despite significant advancements, ABSA research faces three key limitations: (i) a limited focus on quadruple extraction/prediction compared with simpler pair or triplet extraction/prediction tasks, with tasks involving quadruples only recently gaining traction via generative models (Hua et al. 2024); (ii) an insufficient exploration of multimodal ABSA (MABSA) relative to text-based ABSA (Zhao et al. 2024); and (iii) an overemphasis on explicit sentiment analysis compared with implicit sentiment analysis (Zhang et al. 2024). Research addressing implicit ABSA, where aspects and/or opinions lack an explicit textual span, often reduces these implicit elements to uninformative "NULL" labels, such as "apps-Software-NULL-Negative" (Cai, Xia, and Yu 2021). For a detailed discussion of the limitations of ABSA annotation schemes, see Ocampo Diaz, Zhang, and Ng (2020).

To address these three gaps, we adapt the quadruple (*target-aspect-opinion-sentiment*) schema proposed by Li et al. (2023) to multimodal memes and leverage VLMs to recover implicit elements that emerge only through the joint interpretation of text and images. These outputs enable hierarchical clustering and meme corpus analysis.

Meme Clustering

Clustering techniques have been explored to group aspect phrases, including both implicit and explicit aspects (Vargas and Pardo 2018), as well as multilingual aspects (Pessutto, Vargas, and Moreira 2020). In this work, we hierarchically cluster target phrases, aspect phrases, sentiments, and opinion phrases inferred from multimodal memes, enabling a comprehensive analysis of meme corpora at varying levels of granularity. This approach provides deeper insight into the themes and sentiments expressed in memes.

A notable study by Chang (2022) performed meme clustering using DeepCluster (Caron et al. 2018) to learn visual embeddings and differentiate between memes created by regular users and state actors. However, this work relied on a single modality, required manual cluster labeling, and involved the arbitrary preselection of the number of clusters. Another work, MemeCap (Hwang and Shwartz 2023), performs clustering through text-based vector representations. MemeCap clusters memes through their captions, ignoring multimodality in its clustering.

Our method addresses these limitations by clustering semantic elements derived from MABSA, including targets, aspects, sentiments, and opinions. The number of clusters emerges dynamically, eliminating the need for predefined

cluster counts, and the clusters are automatically labeled for interpretability. Furthermore, we advance beyond methods like PromptMTopic (Prakash et al. 2023), a prompt-based approach designed to identify and cluster meme topics using LLMs. While PromptMTopic focuses on distilling high-level topics (e.g., "*politics*"), our method provides a hierarchical and fully articulated representation of meme corpora.

Methodology

Figure 1 (see Appendix C) provides an overview of the MemeTT pipeline, which consists of three key modules: *Data Preprocessing*, *Meme Semantic Extraction*, and *Hierarchical Clustering*. Each module is designed to address specific challenges in the analysis of memes, collectively enabling the scalable, automated analysis of meme corpora.

Data Preprocessing

Memes often feature public figures; identifying such figures facilitates a more comprehensive interpretation of their meaning and significance. To this end, our preprocessing pipeline extracts and refines relevant visual information from memes, enriching downstream semantic analysis and narrative generation.

We use the Amazon Rekognition RecognizeCelebrities endpoint[1] to detect and localize celebrity faces in each meme. We retain up to five highest-confidence matches with confidence scores of at least 90% and draw colored bounding boxes around them. For each retained match, we call the GetCelebrityInfo endpoint[2] to retrieve Wikidata links and extract the associated descriptions. This yields annotated memes, along with celebrity names and descriptions, which are used in downstream analysis when available.

Meme Semantic Extraction

This module extracts rich semantic elements from memes for subsequent hierarchical clustering. We use a prompt (see Appendix F) that integrates six critical steps in a single generation: (1) meme description, (2) text-image relationship classification, (3) intertextual references identification, (4) humor techniques recognition, (5) key message inference, and (6) MABSA.

Meme Description. The VLM generates detailed descriptions of each meme using prompts adapted from Singla et al. (2024) and Schuhmann and Bevan (2023). Our prompt also includes instructions to describe each panel individually when processing multi-panel memes.

Text-Image Relationship Classification. The VLM classifies how text and image interact within each meme by applying a taxonomy of multimodal combinations (McCloud 1994), categorizing text-image pairs as interdependent, text-dominant, image-dominant, synonymous, or additive. To guide this process, we adapt a question from Bettin et al. (2023) that encourages explicit reasoning about how the text contributes to the overall message.

Intertextual References Identification. The VLM identifies and explains the intertextual references that are central to memes (Shifman 2013), including cultural and political allusions (Mukhtar et al. 2024), movie stills and video game imagery (Lankshear and Knobel 2019), and photographs of public figures and events (Polách 2015).

Humor Techniques Recognition. The VLM detects humor techniques by drawing on established meme-humor taxonomies (Schumacher 2024; Catanescu and Tom 2001; Taecharungroj and Nueangjamnong 2015; Buijzen and Valkenburg 2004) and explains how they shape meaning.

Key Message Inference. The VLM infers the key intended messages of each meme within its specific political, social, or cultural context, reflecting its stance.

Multimodal Aspect-Based Sentiment Analysis. This final step involves inferring one or more sets of targets, aspects, opinions, and sentiments expressed in the meme. We also draw on established definitions of "target" (Alaei et al. 2023; Mohammad et al. 2016; Pontiki et al. 2015), "aspect" (Fuyao et al. 2023; Ayub et al. 2022), and "opinion" (Huang et al. 2024) to ensure a nuanced understanding of these elements. This structured approach captures the complex semantic relationships in memes, laying the groundwork for fine-grained clustering and analysis. Extending to MABSA the approach of Zhang et al. (2021), which formulates ABSA as a paraphrasing task, we call on the VLM to rephrase the coarse-grained viewpoints stated in the previous step as one or more structured viewpoints in the following format:

> *The meme views* {*target noun or noun phrase*} *with a* {*sentiment*} *sentiment because its/his/her/their* {*aspect noun or noun phrase*} *is/are seen as* {*opinion adjective or adjective phrase*}.

After inference, if the VLM output contains exactly one identifiable viewpoint, that viewpoint is extracted directly using regular expressions. When the output includes multiple viewpoints or fails to adhere to the expected format, an LLM is prompted to infer and extract a consolidated set of non-redundant viewpoints. A secondary, more flexible regular expression is then applied to extract these viewpoints. Finally, redundant prefixes are removed. This hybrid strategy addresses instances in which the VLM either identifies multiple viewpoints or introduces a viewpoint and subsequently revises or elaborates upon it within the same output.

Hierarchical Clustering

Hierarchical clustering organizes semantic elements (*targets*, *aspects*, *sentiments*, and *opinions*) into multi-level clusters, enabling fine-grained and comprehensive analysis of meme corpora. For instance, at the target level, memes about "*COVID-19 vaccines*" may form a top-level cluster. At the aspect level, these memes may be further grouped into themes such as "*vaccine efficacy and effectiveness*," "*public perception*," and "*safety profile*." Each aspect cluster may

[1]https://docs.aws.amazon.com/rekognition/latest/APIReference/API_RecognizeCelebrities.html

[2]https://docs.aws.amazon.com/rekognition/latest/APIReference/API_GetCelebrityInfo.html

then be partitioned by sentiment polarity: positive, neutral, or negative. At the most granular opinion level within each sentiment cluster, sub-clusters may capture specific orientations, such as "*pandemic-ending*". This hierarchical structure captures the diversity of viewpoints present in meme discourse and facilitates valuable insights into thematic and sentiment trends across the corpus.

The hierarchical clustering of the quadruples comprises *(1) Text Embedding*, *(2) Cluster Assignment*, *(3) Membership Validation*, *(4) Theme Consolidation*, and *(5) Narrative Construction*. Following the generation of text embeddings for targets, aspects, and opinions, the targets undergo *(2) Cluster Assignment* and *(3) Membership Validation*. Then, within each resulting target cluster, aspects are processed through the full clustering sequence: *(2) Cluster Assignment*, *(3) Membership Validation*, and *(4) Theme Consolidation*. Next, these aspect clusters are simply partitioned by their categorical sentiment: positive, neutral, and negative. Subsequently, within each unique (target, aspect, sentiment) cluster, opinions are clustered using the same full sequence (Steps 2-4). At the target level, Step 4 is omitted because targets are open-class and highly variable. Hence, consolidation risks collapsing distinct entities. In contrast, aspects and opinions occupy a more bounded space and may benefit from consolidation. The final step is *Narrative Construction*.

Text Embedding. A high-performance text embedding model is used to encode the extracted *target*, *aspect*, and *opinion* phrases into vector representations.

Cluster Assignment. The text embeddings are then passed to the FINCH algorithm (Sarfraz, Sharma, and Stiefelhagen 2019), which forms the initial clusters based on nearest-neighbor principles. FINCH is chosen for two reasons: it is parameter-free, eliminating the need to predefine the number of clusters, and it is computationally efficient, making it well-suited for clustering high-dimensional embeddings. While methods such as DBSCAN preclude the need for a predefined number of clusters, they remain sensitive to hyperparameter tuning and often require substantial empirical optimization (Huang et al. 2025). In all cases, we retain the first partition returned by FINCH, as it provides the most fine-grained and coherent groupings. An LLM is then used to generate a concise thematic label for each cluster, which yields an initial set of candidate themes.

Membership Validation. At each semantic level (target, aspect, and opinion), an LLM validates the candidate cluster label against the corresponding original element in the quadruple. If the candidate fails this validation, the original pre-clustering text is restored as the label for that instance. Otherwise, the generated label is retained.

Theme Consolidation. For aspects and opinions, we further refine and organize this combined set of retained and restored labels using a structured grouping and labeling approach. Drawing inspiration from the topic collapsing method of PromptMTopic (Prakash et al. 2023) and the topic merging strategy by Pham et al. (2024), this step consolidates synonymous labels, or labels where one clearly encompasses others, into cohesive groups using a prompt-based approach. Each group is assigned a unified label where appropriate, while labels that cannot be meaningfully consolidated remain independent. This resolves semantic overlap, yielding a distinct and non-redundant set of cluster labels at the aspect and opinion levels.

Narrative Construction. Finally, the target, aspect, sentiment, and opinion cluster labels are assembled into concise narrative statements that articulate the viewpoint represented by each meme cluster. These narratives follow the format:

> *The meme(s) view* {*target cluster label*} *with a* {*sentiment*} *sentiment because its/his/her/their* {*aspect cluster label*} *is/are seen as* {*opinion cluster label*}.

These narratives enable researchers and analysts to quickly understand underlying concerns, assess public sentiment, and identify emerging trends.

MemeTT's Models

The MemeTT pipeline uses a combination of VLMs, LLMs, and an embedding model to achieve multimodal reasoning and semantic understanding across diverse meme datasets, and to enable tasks like MABSA and hierarchical clustering.

Vision-Language Models. The VLMs are used in the *Meme Semantic Extraction* module to analyze visual and textual cues from memes. We evaluated several proprietary and open-source VLMs, selecting a diverse range of models guided by the OpenCompass Multimodal Leaderboard benchmark (OpenCompass Contributors 2023).

For proprietary models, *Gemini 2.5 Pro* (gemini-2.5-pro-exp-03-25), *Gemini 2.0 Flash* (gemini-2.0-flash-001), and *Gemini 2.0 Flash Thinking* (gemini-2.0-flash-thinking-exp-01-21) were evaluated (Comanici et al. 2025). Evaluated open-weight models include *Pixtral Large* (pixtral-large-2411), *Mistral Small* (mistral-small-2503), *Pixtral Small* (pixtral-12b-2409) (Agrawal et al. 2024), *LLaMA 4 Maverick* (llama4-maverick-instruct-basic), *LLaMA 4 Scout* (llama4-scout-instruct-basic), *Qwen2.5 VL 32B* (qwen2p5-vl-32b-instruct) (Bai et al. 2025), *Qwen2.5 VL 7B* (Qwen2.5-VL-7B-Instruct) (Bai et al. 2025), *InternVL2.5 8B MPO* (Wang et al. 2025), and *InternVL3 8B* (Zhu et al. 2025). The Mistral and Pixtral models were accessed via the official Mistral AI API, while the larger open-weight models were served using the Fireworks AI API. The remaining models were deployed offline using publicly available weights from Hugging Face.

Large Language Models. Two text-only LLMs were selected to support the downstream, text-based stages of the pipeline. For viewpoint extraction, the *deepseek-r1-basic* model was employed due to its strong performance on the CompassBench LLM Leaderboard (OpenCompass Contributors 2023), particularly in reasoning and instruction-following tasks. For theme generation, membership validation, and theme consolidation, the *qwen3-235b-a22b* (Yang et al. 2025) model was used. This model offers strong performance comparable to *deepseek-r1-basic* (Guo et al. 2025)

while providing improved cost-efficiency for high-volume inference tasks. Both models were served using the Fireworks AI API.

Embedding Model. To generate embeddings for targets, aspects, and opinions, Google's *text-embedding-005* was used. This model has demonstrated strong performance on the MTEB benchmark (Muennighoff et al. 2023) and is well-suited for capturing fine-grained semantic relationships.

Experimental Setup

Datasets

To evaluate the performance and versatility of the MemeTT pipeline, we applied it to a combined pool of three publicly available datasets, namely, TDMeme (Prakash, Hee, and Lee 2023), Harm-C (Pramanick et al. 2021a), and Harm-P (Pramanick et al. 2021b). These datasets were chosen to ensure diversity in topics, allowing for a comprehensive assessment of the pipeline's capabilities. The Harm-C dataset contains 3,544 memes related to the COVID-19 pandemic, providing a challenging testbed for analyzing memes with sensitive and potentially controversial content. The Harm-P dataset focuses on U.S. political discourse, comprising 3,470 memes that reflect diverse opinions and sentiments during politically charged events. The TDMeme dataset comprises 1,876 Singapore-related memes, particularly those associated with Singapore's Total Defence strategy.

In the experiments, the three datasets were combined into a single corpus comprising 8,890 memes. This configuration presents a more challenging setting, designed to evaluate the pipeline's ability to extract meaningful clusters from a diverse and heterogeneous collection. Ideally, the clusters should exhibit strong separation, with the majority of memes within each cluster expressing the same viewpoint.

Implementation Details

All experiments were conducted on Lightning AI Studio. Hardware specifications, computation time, token usage, and model-specific settings are provided in Appendix E. Inference settings were standardized where possible, with a temperature of 0, a *top-K* of 1, and a random seed of 42. While a maximum token limit of 8,192 was used for MABSA, higher token limits were used to allow for more reasoning tokens in other tasks, namely, 16,384 for clustering, 32,768 for LLM-Judge evaluations in RQ1, RQ3, and RQ4, as well as for viewpoint extraction from VLM outputs, and 128,000 for RQ2 to enable reasoning over entire clusters of meme images. For MABSA, invalid outputs, such as cases in which no viewpoint was identified because the output degenerated into repetitive text, were handled using a fallback configuration with a temperature of 1 and, where supported, a *top-K* of 40. Such cases were rare. For clustering, the initial decoding configuration used a temperature of 0 and a *top-K* of 1. If generation did not terminate normally, the same fallback configuration was applied.

Evaluation Overview

The absence of ground-truth semantic quadruples in the datasets makes automated evaluation methods infeasible for this study. As a result, human and LLM-as-a-Judge (LLM-Judge) evaluations were carried out to assess the quality of semantic quadruples generated by the MemeTT pipeline.

Human Evaluation. English-speaking participants were screened for familiarity with the main topics represented in the datasets, including U.S. politics, COVID-19, and Singaporean culture and Total Defence. Owing to the impracticality of recruiting human evaluators well-versed in all three domains, the human evaluation samples were made domain-specific, and vetted evaluators were assigned strictly to the data segments matching their verified domain familiarity. A preliminary screening test, consisting of multiple-choice questions on these topics, was administered to ensure a baseline level of topical understanding. To align evaluators with task expectations, trial runs were conducted prior to the main evaluation. Control questions were also embedded in the main evaluation to ensure reliability. To prevent trivializing the tasks, samples with duplicate memes were manually replaced. This cleaned evaluation set is available in our repository. Nine participants completed the evaluation for RQ1, nine completed RQ3, and eight completed RQ4.

LLM-as-a-Judge Evaluation. For the LLM-Judge evaluation, *GPT-5 mini* (gpt-5-mini-2025-08-07) was used because of its competitiveness on the OpenCompass Multimodal Leaderboard. To ensure robustness against potential response instability in *GPT-5 mini*, whose temperature is fixed at 1, we repeated the LLM-Judge evaluations five times for RQ1, RQ3, and RQ4, and three times for RQ2 to manage the evaluation cost of the numerous baselines involved.

Research Questions

To assess the effectiveness of the pipeline in forming distinct clusters and generating meaningful narratives, we formulate and evaluate four research questions (RQs):

- **RQ1:** Does the *MABSA quadruple* accurately capture the intended message of a meme?
- **RQ2:** Does MemeTT's *hierarchical clustering* produce more homogeneous clusters than *simple clustering*?
- **RQ3:** Do the *clusters* demonstrate strong *intra-cluster quality*, with the narratives accurately capturing the shared semantic elements of most memes within them?
- **RQ4:** Are the *clusters* clearly distinct and easily differentiable from one another (*inter-cluster quality*)?

Experiments and Results

RQ1. Meme Aspect-Based Sentiment Analysis

Experiment Design. The accuracy of each MABSA element (target, aspect, opinion, and sentiment) was evaluated against the intended message of each meme. From each original dataset, 100 unique memes were randomly sampled, together with their corresponding quadruples previously generated by the 12 VLMs. When a VLM generated multiple quadruples for a meme, one quadruple was randomly selected for evaluation. This yielded an evaluation pool of 1,200 quadruples per dataset. To mitigate positional bias, the order of the quadruples was shuffled prior to evaluation.

Model	Target		Aspect		Opinion		Sentiment	
	Human	LLM-Judge	Human	LLM-Judge	Human	LLM-Judge	Human	LLM-Judge
Harm-C								
- Gemini 2.5 Pro	85 / 63	99 / 99	**83 / 56**	97 / 94	**78 / 46**	94 / 89	71 / 38	77 / 68
- Gemini 2.0 Flash Thinking	**89 / 71**	100 / 99	81 / 54	99 / 98	74 / 43	94 / 85	**72 / 39**	85 / 76
- Gemini 2.0 Flash	85 / 58	100 / 97	69 / 45	97 / 90	65 / 35	89 / 75	63 / 33	83 / 68
- LLaMA 4 Maverick	84 / 63	95 / 95	73 / 43	94 / 91	62 / 34	89 / 81	62 / 32	87 / 77
- LLaMA 4 Scout	84 / 57	97 / 92	74 / 40	92 / 82	64 / 28	84 / 72	62 / 27	78 / 68
- Pixtral Large	70 / 39	93 / 89	59 / 27	89 / 82	52 / 22	83 / 73	44 / 19	67 / 59
- Mistral Small	75 / 56	92 / 88	60 / 43	80 / 73	52 / 32	67 / 56	52 / 32	62 / 52
- Pixtral Small	74 / 48	92 / 91	55 / 33	79 / 76	46 / 27	66 / 62	44 / 26	61 / 52
- Qwen2.5 VL 32B	77 / 50	100 / 97	62 / 35	96 / 92	52 / 25	87 / 76	49 / 23	81 / 68
- Qwen2.5 VL 7B	74 / 46	98 / 96	53 / 25	90 / 83	40 / 15	70 / 62	38 / 15	55 / 43
- InternVL3 8B	**89** / 54	99 / 95	65 / 31	96 / 86	50 / 20	79 / 65	47 / 19	75 / 60
- InternVL2.5 8B MPO	77 / 52	98 / 96	62 / 35	94 / 87	51 / 27	75 / 62	38 / 22	52 / 42
Harm-P								
- Gemini 2.5 Pro	**88 / 71**	100 / 100	**80 / 56**	100 / 99	**78 / 47**	97 / 95	72 / **45**	88 / 85
- Gemini 2.0 Flash Thinking	87 / **71**	98 / 97	78 / 55	97 / 95	72 / 45	95 / 91	68 / 43	90 / 82
- Gemini 2.0 Flash	86 / **71**	98 / 98	**80 / 56**	96 / 94	75 / 45	92 / 87	**73** / 40	90 / 83
- LLaMA 4 Maverick	80 / 57	97 / 94	70 / 39	95 / 85	64 / 34	92 / 81	64 / 33	90 / 76
- LLaMA 4 Scout	80 / 62	99 / 99	71 / 34	94 / 89	67 / 27	83 / 77	63 / 27	79 / 73
- Pixtral Large	76 / 57	95 / 93	69 / 42	86 / 83	56 / 37	81 / 74	54 / 36	72 / 66
- Mistral Small	85 / 64	92 / 90	77 / 46	87 / 77	69 / 35	78 / 68	65 / 31	74 / 65
- Pixtral Small	70 / 45	92 / 90	58 / 32	85 / 81	50 / 25	75 / 67	49 / 25	69 / 58
- Qwen2.5 VL 32B	75 / 56	97 / 96	64 / 38	87 / 85	57 / 33	78 / 73	54 / 32	71 / 65
- Qwen2.5 VL 7B	68 / 53	93 / 89	51 / 29	79 / 72	43 / 15	67 / 60	42 / 15	59 / 54
- InternVL3 8B	73 / 51	97 / 94	59 / 29	83 / 75	46 / 20	73 / 59	45 / 19	66 / 52
- InternVL2.5 8B MPO	76 / 61	98 / 96	59 / 32	93 / 88	51 / 22	78 / 65	43 / 20	60 / 48
TDMeme								
- Gemini 2.5 Pro	**89 / 67**	100 / 100	**80 / 55**	99 / 98	**75 / 49**	97 / 94	**69 / 40**	73 / 69
- Gemini 2.0 Flash Thinking	78 / 48	97 / 97	61 / 31	95 / 94	49 / 25	91 / 88	44 / 21	82 / 76
- Gemini 2.0 Flash	71 / 45	99 / 96	57 / 29	99 / 90	48 / 22	90 / 81	44 / 19	85 / 74
- LLaMA 4 Maverick	67 / 41	98 / 95	49 / 27	92 / 87	39 / 22	85 / 79	33 / 20	78 / 71
- LLaMA 4 Scout	53 / 24	95 / 95	38 / 11	90 / 86	28 / 10	84 / 74	25 / 10	78 / 64
- Pixtral Large	53 / 30	96 / 91	36 / 19	91 / 82	31 / 14	81 / 71	20 / 9	57 / 46
- Mistral Small	49 / 27	94 / 88	35 / 16	81 / 72	28 / 14	70 / 63	26 / 10	62 / 54
- Pixtral Small	47 / 25	91 / 86	30 / 14	80 / 74	23 / 10	66 / 56	18 / 7	53 / 45
- Qwen2.5 VL 32B	54 / 24	97 / 95	39 / 15	94 / 88	30 / 11	89 / 74	20 / 10	74 / 61
- Qwen2.5 VL 7B	46 / 22	90 / 88	25 / 9	80 / 71	16 / 5	60 / 50	12 / 5	49 / 37
- InternVL3 8B	45 / 23	93 / 88	29 / 8	84 / 77	22 / 4	68 / 58	16 / 3	60 / 48
- InternVL2.5 8B MPO	45 / 19	91 / 88	27 / 8	81 / 79	18 / 6	62 / 53	17 / 6	48 / 43

Table 1: MABSA results based on 100 sampled memes per dataset. Human = Number of instances that human evaluators judged to be accurate (majority / unanimous vote). LLM-Judge = Number of instances that GPT-5 mini judged to be accurate (majority / unanimous vote). The best and second-best human evaluation results are indicated in bold and underlined, respectively.

For human evaluation, annotators reviewed the semantic elements of each quadruple sequentially, beginning with the target, followed by the aspect, opinion, and sentiment. The evaluation process was terminated when annotators encountered the first incorrect semantic element in the sequence, to reduce unnecessary cognitive load. Annotators were encouraged to consult external sources to clarify unfamiliar terms or verify the identities of individuals depicted in the memes. For the LLM-Judge approach, each quadruple was assessed in the same sequential manner.

Experiment Results. Table 1 reports accuracy scores for human evaluators (three evaluators per dataset) and the LLM-Judge (five evaluation runs), based on both majority and unanimous agreement. Overall, *Gemini 2.5 Pro* emerges as the best-performing model for MABSA generation. Under human evaluation, it outperforms most other models, achieving the highest accuracy in nine out of twelve cases (i.e., four MABSA elements evaluated across three datasets) under both voting conditions. The LLM-Judge results reveal a similar pattern, with *Gemini 2.5 Pro* ranking at or near the top for target, aspect, and opinion. However, it exhibits a noticeable decline in the LLM-Judge rankings for sentiment on the Harm-C and TDMeme datasets. In addition, both evaluation methods reveal a consistent downward trend in performance across semantic levels. Models show increased dif-

Clustering	Num. C	Macro-Purity	Micro-Purity
Text Embedding			
text-embedding-005 (F)	1701	.680	.537
text-embedding-005 (K)†		.648	.543
embed-v4.0 (F)	1604	.682	.508
embed-v4.0 (K)†		.661	.510
Multimodal Embedding			
voyage-multimodal-3 (F)	1736	.632	.522
voyage-multimodal-3 (K)†		.618	.522
embed-v4.0 (F)	1676	.644	.506
embed-v4.0 (K)†		.627	.514
Viewpoint Embedding			
text-embedding-005 (F)†	2484	.742	.644
text-embedding-005 (K)†		.747	.645
MemeTT			
Gemini 2.5 Pro (F)†	5810	.828	.750

Table 2: Cluster homogeneity evaluation (mean of three LLM-Judge runs). Num. C = the number of clusters. F = FINCH. K = K-means. The dagger sign (†) indicates approaches that yielded singleton clusters, each contributing purity 1. Embedding dimensionalities: Google's text-embedding-005 ($d = 768$), Cohere's embed-v4.0 ($d = 1,536$), Voyage AI's voyage-multimodal-3 ($d = 1,024$).

ficulty in accurately identifying lower-level elements such as aspects, opinions, and sentiments, particularly in the TD-Meme dataset.

Discussion. Although both human and LLM-Judge evaluations exhibit a similar downward trend in performance as granularity increases, the LLM-Judge tends to assess quadruple elements more leniently than human evaluators. For example, on Harm-P, *InternVL3 8B* received sentiment accuracy scores of 66 and 52 from the LLM-Judge under majority and unanimous agreement, compared with only 45 and 19 from human evaluators under the same conditions. This highlights a clear gap between human and LLM assessments (see Appendix D for case studies), pointing to a limitation of the LLM-Judge approach and demonstrating the importance of human involvement for a more comprehensive evaluation. Further analysis suggests that the divergence may be due to inconsistencies in the LLM-Judge assessments. In some cases, different VLMs inferred conflicting opinions or sentiments toward the same target and aspect within a single meme, yet the LLM-Judge considered both interpretations accurate.

RQ2. Hierarchical Clustering

Experiment Design. To evaluate whether MemeTT produces homogeneous clusters, we compared its opinion-level clusters, derived from quadruples generated by *Gemini 2.5 Pro* (the best-performing MABSA model), with those of ten baseline approaches, using an LLM-Judge. By homogeneity, we mean that the memes in a cluster share the same viewpoint, not merely a topic, because topic-level clusters conflate opposing stances and rationales. For a cluster with n unique memes, the judge was shown only the meme images and asked to return the size of the largest subset that shared the same viewpoint, denoted by $m \in \{1, \ldots, n\}$, with $m = 1$ when no two memes matched. Drawing on the concept of cluster purity used in large-scale document clustering (Nayak et al. 2010), but adapted to our setting, which lacks ground-truth categories, we define cluster purity as m/n and aggregate across K clusters. We calculate micro-purity $\left(\sum_{i=1}^{K} m_i / \sum_{i=1}^{K} n_i\right)$ as a weighted average across all cluster assignments, and macro-purity $\left(\frac{1}{K}\sum_{i=1}^{K} m_i/n_i\right)$ as an unweighted average over the K clusters produced. Singleton clusters were assigned a purity of 1.

Baselines. Three types of baselines were considered, yielding ten baselines in total: (1) clustering using meme text only (text embeddings), (2) clustering using multimodal embeddings, and (3) clustering using text embeddings of viewpoints generated by *Gemini 2.5 Pro*. The first two baseline types were naive approaches that did not perform intermediate processing such as MABSA, whereas the third was an ablation of MemeTT that evaluated the effect of removing hierarchical clustering while retaining MABSA-based representations. The meme-text baselines operated on text extracted using the Google Cloud Vision API[3]. Memes with no detectable text were excluded, resulting in 8,882 memes for these baselines, compared with 8,890 in the full corpus used when multimodal embeddings were involved. For MemeTT and the baselines using text embeddings of viewpoints generated by *Gemini 2.5 Pro*, clustering operated on the 12,242 generated viewpoints[4], as a single meme could yield multiple viewpoints. We evaluated different embedding models and clustering algorithms, as shown in Table 2. For all FINCH-based baselines, we used the first partition, as in MemeTT. For all K-means-based baselines, we set the number of clusters equal to that returned by the corresponding FINCH run.

Experiment Results. As reported in Table 2, MemeTT achieves the strongest clustering quality, with a mean macro-purity of 0.828 and mean micro-purity of 0.750 across 5,810 clusters. Compared with the strongest baseline (viewpoint-embedding K-means clustering: 0.747 macro / 0.645 micro across 2,484 clusters), MemeTT improves mean macro-purity by 0.081 and mean micro-purity by 0.105. As a robustness check, we also report purity scores excluding singleton clusters for MemeTT and viewpoint-embedding FINCH clustering, the two best-performing approaches among those that contain singletons. Under this stricter condition, MemeTT achieves a mean macro-purity of 0.698 and mean micro-purity of 0.685, ranking second in macro-purity and first in micro-purity. The strongest baseline under the same exclusion, viewpoint-embedding (FINCH), achieves

[3]https://docs.cloud.google.com/vision/docs/ocr

[4]We identified, post-acceptance, a case-normalization bug that affected the clustering of 5.66% (693 of 12,242) of the quadruples/viewpoints in MemeTT. We recalculated all evaluation metrics after excluding the affected narratives. The overall conclusions remained unchanged. See Appendix I for recalculations.

0.718 macro / 0.637 micro. We note that excluding singletons penalizes methods that produce more singleton clusters. Specifically, MemeTT has 2,501 singleton clusters, compared with 636 for viewpoint-embedding (K-means) and 141–215 for the remaining baselines with singletons. In practice, genuinely unique viewpoints are likely common and operationally important. Assigning singletons a purity score of 1 therefore remains more appropriate.

Discussion. We note that the baselines in Table 2 differ not only in clustering algorithm but also in input representation. Text and multimodal baselines cluster memes directly from their embeddings, whereas viewpoint-embedding clustering and MemeTT derive clusters from MABSA outputs. Moreover, because MABSA can extract multiple viewpoints from a single meme, a meme that praises one subject while denouncing another may appear in more than one cluster in MemeTT and in the viewpoint-embedding baselines. The text and multimodal baselines, however, assign each meme to exactly one cluster, losing information. Accordingly, the results should be interpreted as comparing the purity of the final meme groupings produced by each method.

The higher number of clusters is a direct result of the method's hierarchical approach, which organizes MABSA elements by targets, aspects, sentiments, and opinions. This structure produces many small, precise leaf clusters while preserving large aggregates at upper levels, making triage more practical. Analysts can first prioritize high-volume targets, then drill into key aspects and sentiments, and finally inspect specific opinions. Flat baselines simply cluster memes in a single pass. Their largest clusters contain fewer than 80 items, while MemeTT's largest target-level cluster contains nearly a thousand memes, allowing clearer prioritization before fine-grained review.

RQ3. Intra-Cluster Quality

Experiment Design. We evaluated intra- and inter-cluster quality on clusters derived from quadruples generated by *Gemini 2.5 Pro*, which was selected for its strong performance in the MABSA task. Intra-cluster quality was evaluated by human annotators and the LLM-Judge, who assessed narrative coherence and the relevance of the narratives to the memes in each cluster. To ensure manageability, we randomly sampled clusters of three to five memes.

We drew the evaluation set from clusters whose memes all originated from the same source dataset for two reasons. First, it was difficult to recruit annotators with sufficient domain knowledge covering diverse topics, such as U.S. politics and Singapore's Total Defence. Second, approximately 80% of the quadruples were already grouped into such single-source clusters. Evaluators assessed both narrative coherence and relevance to the quadruple elements reflected in each cluster's memes. Narrative coherence was rated on a four-point Likert scale. To assess relevance, annotators performed a sequential evaluation. They first counted the memes for which the narrative captured the target. Within that subset, they counted the memes for which it also captured the aspect. Within that subset, they counted those for which it captured the opinion and sentiment.

		Coherence					
$\|C\|$	N	Hum.	LLM-J	Agr_h	Agr_{hl}	$AC1_h$	$AC1_{hl}$
Harm-C							
3	50	3.69	3.90	24	24	.576	.655
4	50	3.71	3.86	18	17	.524	.596
5	22	3.77	3.90	10	10	.622	.714
Harm-P							
3	50	3.92	3.95	40	39	.853	.878
4	50	3.93	3.97	42	42	.889	.910
5	41	3.93	3.99	35	35	.899	.921
TDMeme							
3	50	3.63	3.97	16	16	.441	.546
4	39	3.66	3.93	14	13	.500	.577
5	15	3.67	3.91	5	5	.530	.566

Table 3: Cluster narrative coherence evaluation (LLM-Judge values are means of 5 runs). $|C|$ = Number of memes in cluster. N = Number of clusters assessed. Hum. = Average human score. LLM-J = GPT-5 mini (average score). Agr_h = The number of instances in which all human annotators independently assigned the maximum score of 4. Agr_{hl} = The number of instances in which all human annotators and the LLM-Judge (majority vote of 5 runs) independently assigned the maximum score of 4. $AC1_h$ = Gwet's AC1 between human evaluators. $AC1_{hl}$ = Gwet's AC1 treating the majority vote of the LLM-Judge as an additional evaluator.

Experiment Results. Tables 3 and 4 present the results of the human evaluation (three evaluators per dataset) and the LLM-Judge (five evaluation runs) for narrative coherence and the relevance of the narratives to the memes within their corresponding clusters, respectively. Across datasets and cluster sizes, the average coherence scores assigned by human evaluators were consistently close to the maximum score of 4. This suggests that the target, aspect, and opinion cluster labels are well-aligned with one another and that the resulting narratives are fluent and coherent. Inter-annotator agreement for coherence, measured using Gwet's AC1, is also reported in Table 3. We used Gwet's AC1 instead of Krippendorff's alpha because it is more robust to class imbalance, a prevalent issue in our case, since most ratings skewed toward 4. If no strict majority was obtained across the five LLM-Judge runs, the LLM-Judge's rating for that sample was treated as missing in the agreement calculation. The AC1 scores indicate moderate to strong agreement among human evaluators, with Harm-P achieving the highest values ($AC1_h$ = 0.853–0.899), followed by Harm-C (0.524–0.622) and TDMeme (0.441–0.530). Including the majority vote of the LLM-Judge as a fourth evaluator consistently increases agreement ($AC1_{hl} > AC1_h$ in all cases), suggesting that the LLM-Judge is broadly aligned with the human evaluations. For relevance, the sequential evaluation design, where the number of memes counted at the aspect level depends on an annotator's own target-level count and their count at the opinion-sentiment level depends on their

		Target				Aspect				Opinion-Sentiment			
$\|C\|$	N	Hum.	LLM-J	Agr_h	Agr_{hl}	Hum.	LLM-J	Agr_h	Agr_{hl}	Hum.	LLM-J	Agr_h	Agr_{hl}
Harm-C													
3	50	2.81	2.85	34	32	2.53	2.68	20	17	2.28	2.16	11	7
4	50	3.64	3.81	29	28	3.39	3.64	21	19	3.09	3.10	16	11
5	22	4.48	4.75	12	12	4.05	4.53	6	6	3.58	3.85	3	2
Harm-P													
3	50	2.84	2.87	41	41	2.48	2.49	22	18	2.41	2.06	21	13
4	50	3.79	3.74	36	33	3.57	3.22	21	15	3.43	2.56	16	6
5	41	4.78	4.78	24	24	4.39	4.28	12	11	4.14	3.73	11	10
TDMeme													
3	50	2.63	2.68	22	20	2.10	2.17	5	4	1.80	1.85	2	2
4	39	3.62	3.67	21	21	2.89	3.18	3	3	2.34	2.61	0	0
5	15	4.42	4.60	7	6	3.33	3.57	0	0	2.71	2.51	0	0

Table 4: Relevance of cluster narratives to the quadruple elements reflected in each cluster's memes (LLM-Judge values are means of 5 runs). $|C|$ = Number of memes in cluster. N = Number of clusters assessed. Hum. = Average human score. LLM-J = GPT-5 mini (average score). Agr_h = The number of instances in which all human annotators independently assigned the maximum score of $|C|$, Agr_{hl} = The number of instances in which all human annotators and the LLM-Judge (majority vote of 5 runs) independently assigned the maximum score of $|C|$.

aspect-level count, makes standard inter-annotator agreement measures ill-defined at these finer-grained levels.

The relevance scores for each semantic element further indicate that the narratives accurately reflect the majority of memes in each cluster with respect to their targets, aspects, opinions, and sentiments. Narrative relevance is strongest at the target and aspect levels, with a consistent decline at the opinion-sentiment level. This pattern holds across all datasets and cluster sizes, with the weakest performance on the opinion-sentiment level in the TDMeme dataset, where the number of instances in which all human annotators independently assigned the maximum score, Agr_h, drops to 0 for $|C| \geq 4$. While human evaluators assessed narrative relevance by counting matching memes within a cluster, the LLM-Judge evaluated relevance by assessing one narrative against one meme at a time using the same sequential logic. The LLM-Judge results corroborate these findings, indicating high coherence and reasonable relevance overall, with the same relative weakness at the opinion-sentiment level.

Discussion. The average coherence and relevance scores across datasets and cluster sizes indicate that the targets, aspects, and opinions in the narratives fit well together, and that the narratives reflect the salient ideas of most of their constituent memes. TDMeme's weaker relevance performance is expected given its Singapore-specific, niche content, which is likely underrepresented in pretraining corpora. This interpretation aligns with our RQ1 results, as we observed a performance decline across semantic levels in both human and LLM-Judge evaluations, with the most pronounced degradation on TDMeme.

RQ4. Inter-Cluster Quality

Experiment Design. Cluster distinctiveness was evaluated by presenting human annotators and an LLM-Judge with a narrative and three memes, two drawn from the same opinion-level cluster and one from a different cluster. The evaluators were tasked with identifying the meme that was least aligned with the narrative. For human evaluation, we used three annotators per dataset, and a cluster was considered differentiable if at least two annotators correctly selected the meme originating from a different cluster. For the LLM-Judge, we ran the evaluation five times per sample and used the majority-vote prediction as the final label. For agreement analysis, this majority-vote LLM-Judge label was treated as an additional evaluator alongside the three human annotators. If no strict majority was obtained, the LLM-Judge's rating was counted as incorrect for accuracy and as

Dataset	Trg.	Asp.	Snt.	Human	LLM-Judge	α_h	α_{hl}
Harm-C	✗	✗	✗	96.67%	96.67%	.831	.855
	✓	✗	✗	70.00%	70.00%	.620	.549
	✓	✓	✗	63.33%	63.33%	.333	.348
	✓	✓	✓	43.33%	30.00%	.345	.322
Harm-P	✗	✗	✗	93.33%	86.67%	.967	.925
	✓	✗	✗	66.67%	80.00%	.494	.574
	✓	✓	✗	56.67%	83.33%	.389	.398
	✓	✓	✓	44.83%	65.52%	.368	.377
TDMeme	✗	✗	✗	93.33%	96.67%	.718	.750
	✓	✗	✗	50.00%	63.33%	.298	.392
	✓	✓	✗	37.04%	51.85%	.447	.424
	✓	✓	✓	43.33%	70.00%	.517	.443

Table 5: Inter-cluster quality experiment results. Trg. = Target. Asp. = Aspect. Snt. = Sentiment. Human = Human (majority vote). LLM-Judge = GPT-5 mini (majority vote). α_h = Krippendorff's Alpha between human evaluators. α_{hl} = Krippendorff's Alpha treating the majority vote of the LLM-Judge as an additional evaluator.

missing in the agreement calculation. This evaluation was carried out under four increasingly difficult scenarios: (1) the most difficult setting, where the memes shared the same target, aspect, and sentiment clusters but differed in opinion; (2) a highly difficult setting, where the memes shared the same target and aspect but differed in sentiment; (3) a moderate setting, where the memes shared the same target but differed in aspect; and (4) the least difficult setting, where the memes originated from entirely different target clusters.

For each dataset and difficulty setting, we sampled 30 clusters and drew two memes per cluster exclusively from that dataset. In nearly all sampled clusters, that dataset was the majority source. To preserve both the integrity and the difficulty of the evaluation, the third meme, drawn from a different cluster, was also selected from the same dataset while satisfying the specific conditions of the evaluation scenario. Three TDMeme samples from Setting 2 (highly difficult) and one Harm-P sample from Setting 1 (most difficult) were removed from the evaluation pool after we identified a sampling bug that caused all three memes in each of these samples to originate from the same cluster. As a result, Setting 2 for TDMeme includes 27 samples and Setting 1 for Harm-P includes 29 samples, rather than 30.

Results. Table 5 reports the accuracy of identifying imposter memes across clusters under varying levels of difficulty, alongside Krippendorff's alpha scores measuring inter-evaluator agreement. Both human and LLM-Judge evaluation results indicate that distinguishing imposter memes becomes increasingly difficult when clusters share overlapping semantic dimensions, such as targets and aspects. At the target level, separation is straightforward. When the imposter meme differs in its target, accuracies for human evaluators and the LLM-Judge are at least 93.33% and 86.67%, respectively, with high agreement. At the opinion level, where the task is to distinguish memes that share the same target, aspect, and sentiment but differ only in opinion, accuracies decline to as low as 43.33% and 30.00% for human evaluators and the LLM-Judge, respectively, with agreement declining to the 0.3 range for Harm-C and Harm-P.

Discussion. These patterns suggest that MemeTT produces well-separated clusters at a coarse granularity, but separability weakens when narratives share similar semantic elements of the target and aspect, leaving only finer cues, such as sentiment and opinion, to distinguish clusters. This trend may reflect the comparatively lower extraction accuracy of these underlying semantic elements at finer-grained levels, as observed in RQ1.

Conclusion

In this work, we presented MemeTT, the first framework, to the best of our knowledge, that integrates VLMs, LLMs, and MABSA to hierarchically cluster memes based on their targets, aspects, sentiments, and opinions. Our human evaluation experiments demonstrated that the best-performing models excelled at identifying high-level semantic elements, such as targets and aspects, but were challenged when inferring lower-level elements, such as opinions and sentiments. The evaluation of the clusters revealed that they were homogeneous and that the narratives were relevant to the majority of their constituent memes, although there remains room for improvement in the distinctiveness of the clusters at the lower sentiment and opinion levels. This could be overcome as techniques in MABSA improve and with the curation of a gold-standard MABSA dataset for memes, which, to the best of our knowledge, is currently non-existent.

References

Agrawal, P.; Antoniak, S.; Hanna, E. B.; et al. 2024. Pixtral 12B. *CoRR*, abs/2410.07073.

Al Rousan, R.; Al Harahsheh, A.; and Al Rousan, S. 2023. "What's in a meme?" A thematic analysis of memes related to COVID-19 in Jordan. *Onomázein*, 61: 212–235.

Alaei, A.; Wang, Y.; Bui, V.; and Stantic, B. 2023. Target-Oriented Data Annotation for Emotion and Sentiment Analysis in Tourism Related Social Media Data. *Future Internet*, 15(4).

Ayub, N.; Talib, M. R.; Hanif, M. K.; and Awais, M. 2022. Aspect Extraction Approach for Sentiment Analysis Using Keywords. *Computers, Materials and Continua*, 74(3): 6879–6892.

Bai, S.; Chen, K.; Liu, X.; et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bettin, B.; Sarabia, A.; Gonzalez, M. C.; Gatti, I.; Magnan, C.; Murav, N.; Vanden Heuvel, R.; McBride, D.; and Abraham, S. 2023. Say What You Meme: Exploring Memetic Comprehension Among Students and Potential Value of Memes for CS Education Contexts. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1*, ICER '23, 416–429. New York, NY, USA: Association for Computing Machinery. ISBN 9781450399760.

Bijker, R.; Merkouris, S. S.; Dowling, N. A.; and Rodda, S. N. 2024. ChatGPT for Automated Qualitative Research: Content Analysis. *J Med Internet Res*, 26: e59050.

Buijzen, M.; and Valkenburg, P. M. 2004. Developing a Typology of Humor in Audiovisual Media. *Media Psychology*, 6(2): 147–167.

Cai, H.; Xia, R.; and Yu, J. 2021. Aspect-Category-Opinion-Sentiment Quadruple Extraction with Implicit Aspects and Opinions. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 340–350. Online: Association for Computational Linguistics.

Cao, R.; Hee, M. S.; Kuek, A.; Chong, W.-H.; Lee, R. K.-W.; and Jiang, J. 2023. Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 5244–5252. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep Clustering for Unsupervised Learning of Visual Features. In *Computer Vision – ECCV 2018: 15th European*

Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV, 139–156. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-01263-2.

Catanescu, C.; and Tom, G. 2001. Types of Humor in Television and Magazine Advertising. *Review of business*, 22(1): 92.

Chang, K.-C. 2022. Mapping Visual Themes among Authentic and Coordinated Memes.

Cho, J. Y.; and Lee, E.-H. 2014. Reducing Confusion about Grounded Theory and Qualitative Content Analysis: Similarities and Differences. *The Qualitative Report*, 1–20.

Comanici, G.; Bieber, E.; Schaekermann, M.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Fuyao, Z.; Yijia, Z.; Mengyi, W.; Hong, Y.; Mingyu, L.; and Liang, Y. 2023. Self Question-answering: Aspect Sentiment Triplet Extraction via a Multi-MRC Framework based on Rethink Mechanism. In Sun, M.; Qin, B.; Qiu, X.; Jiang, J.; and Han, X., eds., *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, 701–712. Harbin, China: Chinese Information Processing Society of China.

Guo, D.; Yang, D.; Zhang, H.; et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081): 633–638.

Hee, M. S.; Chong, W.-H.; and Lee, R. K.-W. 2023. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23. ISBN 978-1-956792-03-4.

Hua, Y. C.; Denny, P.; Wicker, J.; and Taskova, K. 2024. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artif. Intell. Rev.*, 57(11): 296.

Huang, P.; Xiao, X.; Xu, Y.; and Chen, J. 2024. DMIN: A Discourse-specific Multi-granularity Integration Network for Conversational Aspect-based Sentiment Quadruple Analysis. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 16326–16338. Bangkok, Thailand: Association for Computational Linguistics.

Huang, Z.; Liang, Z.; Zhou, S.; and Zhang, S. 2025. An Improved Density-Based Spatial Clustering of Applications with Noise Algorithm with an Adaptive Parameter Based on the Sparrow Search Algorithm. *Algorithms*, 18(5).

Hwang, E.; and Shwartz, V. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1433–1445. Singapore: Association for Computational Linguistics.

Instagram. 2020. Instagram Year in Review: How Memes Were the Mood of 2020. https://about.instagram.com/blog/announcements/instagram-year-in-review-how-memes-were-the-mood-of-2020. Accessed: 2026-03-17.

Johann, M. 2022. Political participation in transition: Internet memes as a form of political expression in social media. *Studies in Communication Sciences*, 22(1): 149–164.

Lankshear, C.; and Knobel, M. 2019. Memes, Macros, Meaning, and Menace: Some Trends in Internet Memes. *The journal of communication and media studies (Champaign)*, 4(4): 43–57.

Li, B.; Fei, H.; Li, F.; Wu, Y.; Zhang, J.; Wu, S.; Li, J.; Liu, Y.; Liao, L.; Chua, T.-S.; and Ji, D. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 13449–13467. Toronto, Canada: Association for Computational Linguistics.

Lim, Y. Y.; Hee, M. S.; Yee, X. W.; Yau, W. K.; Sim, X.; Tay, W.; Ng, W. S.; Ng, S.-K.; and Lee, R. K.-W. 2024. AISG's Online Safety Prize Challenge: Detecting Harmful Social Bias in Multimodal Memes. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, 1884–1891. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701726.

McCloud, S. 1994. *Understanding Comics: The Invisible Art*. New York: Harper Perennial, 1st harperperennial edition. ISBN 9780060976255;006097625X.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In Bethard, S.; Carpuat, M.; Cer, D.; Jurgens, D.; Nakov, P.; and Zesch, T., eds., *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. San Diego, California: Association for Computational Linguistics.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2014–2037. Dubrovnik, Croatia: Association for Computational Linguistics.

Mukhtar, S.; Ayyaz, Q. U. A.; Khan, D. S.; Bhopali, A. M. N.; Saifullah; Sajid, D. M. K. M.; and Babbar, D. A. W. 2024. Memes In The Digital Age: A Sociolinguistic Examination Of Cultural Expressions And Communicative Practices Across Border. *Educational Administration: Theory and Practice*, 30(6): 1443–1455.

Nayak, R.; De Vries, C. M.; Kutty, S.; Geva, S.; Denoyer, L.; and Gallinari, P. 2010. Overview of the INEX 2009 XML Mining Track: Clustering and Classification of XML Documents. In Geva, S.; Kamps, J.; and Trotman, A., eds., *Focused Retrieval and Evaluation*, 366–378. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-14556-8.

Ocampo Diaz, G.; Zhang, X.; and Ng, V. 2020. Aspect-Based Sentiment Analysis as Fine-Grained Opinion Mining. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and*

Evaluation Conference, 6804–6811. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

OpenCompass Contributors. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models. https://github.com/open-compass/opencompass.

Pessutto, L. R. C.; Vargas, D. S.; and Moreira, V. P. 2020. Multilingual aspect clustering for sentiment analysis. *Knowledge-Based Systems*, 192: 105339.

Pham, C. M.; Hoyle, A.; Sun, S.; Resnik, P.; and Iyyer, M. 2024. TopicGPT: A Prompt-based Topic Modeling Framework. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2956–2984. Mexico City, Mexico: Association for Computational Linguistics.

Polách, V. P. 2015. Memes, Trojan Horses and the Discursive Power of Audience. *Human Affairs*, 25(2): 189–203.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In Nakov, P.; Zesch, T.; Cer, D.; and Jurgens, D., eds., *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 486–495. Denver, Colorado: Association for Computational Linguistics.

Prakash, N.; Hee, M. S.; and Lee, R. K.-W. 2023. TotalDefMeme: A Multi-Attribute Meme dataset on Total Defence in Singapore. In *Proceedings of the 14th ACM Multimedia Systems Conference*, MMSys '23, 369–375. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701481.

Prakash, N.; Wang, H.; Hoang, N. K.; Hee, M. S.; and Lee, R. K.-W. 2023. PromptMTopic: Unsupervised Multimodal Topic Modeling of Memes using Large Language Models. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 621–631. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Pramanick, S.; Dimitrov, D.; Mukherjee, R.; Sharma, S.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021a. Detecting Harmful Memes and Their Targets. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2783–2796. Online: Association for Computational Linguistics.

Pramanick, S.; Sharma, S.; Dimitrov, D.; Akhtar, M. S.; Nakov, P.; and Chakraborty, T. 2021b. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4439–4455. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Sarfraz, S.; Sharma, V.; and Stiefelhagen, R. 2019. Efficient Parameter-Free Clustering Using First Neighbor Relations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8926–8935.

Schuhmann, C.; and Bevan, P. 2023. LAION POP: 600,000 HIGH-RESOLUTION IMAGES WITH DETAILED DESCRIPTIONS. https://laion.ai/blog/laion-pop/. Accessed: 2024-10-12.

Schumacher, D. 2024. *Political humor on TikTok: A mixed methods approach to the use of humor by creators during the Dutch parliamentary elections*. Master's thesis, Erasmus School of History, Culture and Communication.

Shifman, L. 2013. *Memes in Digital Culture*. The MIT Press. ISBN 9780262317696.

Singla, V.; Yue, K.; Paul, S.; Shirkavand, R.; Jayawardhana, M.; Ganjdanesh, A.; Huang, H.; Bhatele, A.; Somepalli, G.; and Goldstein, T. 2024. From Pixels to Prose: A Large Dataset of Dense Image Captions. *CoRR*, abs/2406.10328.

Taecharungroj, V.; and Nueangjamnong, P. 2015. Humour 2.0: Styles and Types of Humour and Virality of Memes on Facebook. *Journal of Creative Communications*, 10(3): 288–302.

Toh, S.; Kuek, A.; Chong, W.-H.; and Lee, R. K.-W. 2023. MERMAID: A Dataset and Framework for Multimodal Meme Semantic Understanding. In *2023 IEEE International Conference on Big Data (BigData)*, 433–442.

Vargas, F. A.; and Pardo, T. A. S. 2018. Aspect Clustering Methods for Sentiment Analysis. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, 365–374. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-319-99721-6.

Vears, D. F.; and Gillam, L. 2022. Inductive content analysis: A guide for beginning qualitative researchers. *Focus on Health Professional Education: A Multi-Professional Journal*, 23(1): 111–127.

Wang, W.; Chen, Z.; Wang, W.; et al. 2025. Enhancing the Reasoning Ability of Multimodal Large Language Models via Mixed Preference Optimization. arXiv:2411.10442.

Yang, A.; Li, A.; Yang, B.; et al. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yus, F. 2019. *Multimodality in Memes: A Cyberpragmatic Approach*, 105–131. Cham: Springer International Publishing. ISBN 978-3-319-92663-6.

Zhang, H.; Cheah, Y.; Alyasiri, O. M.; and An, J. 2024. Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: a comprehensive survey. *Artif. Intell. Rev.*, 57(2): 17.

Zhang, W.; Deng, Y.; Li, X.; Yuan, Y.; Bing, L.; and Lam, W. 2021. Aspect Sentiment Quad Prediction as Paraphrase Generation. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9209–9219. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zhao, H.; Yang, M.; Bai, X.; and Liu, H. 2024. A Survey on Multimodal Aspect-Based Sentiment Analysis. *IEEE Access*, 12: 12039–12052.

Zhu, J.; Wang, W.; Chen, Z.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, the paper advances automatic sentiment analysis of existing vision-language memes widespread on the Internet, which does not violate any social contracts.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. They are well-defined in the Abstract and Introduction.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. They are defined in the introductory paragraph after Methodology.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. They are described under the "Datasets" in the "Experimental Setup" section.
   (e) Did you describe the limitations of your work? Yes, please refer to the "Appendix A - Limitations".
   (f) Did you discuss any potential negative societal impacts of your work? Yes. Refer to "Appendix B - Ethics Statement" section.
   (g) Did you discuss any potential misuse of your work? Yes. Refer to "Appendix B - Ethics Statement" section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we used greedy decoding to ensure reproducibility and provided the prompts in Appendix F.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. The code, data and instructions are submitted as supplemental materials. We will upload to GitHub upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. These details can be found under the "Experimental Setup" section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A. We did not train any model, and we used greedy decoding for LLM inference for reproducibility.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes. Refer to "Appendix E - Computational Cost and Settings" section.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. Refer to "Experiments and Results" section.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. Refer to "Appendix B - Ethics Statement" section.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes. Refer to the "Experimental Setup" section.
   (b) Did you mention the license of the assets? N/A. We are using open-source dataset assets permitted for research, and are not releasing any new datasets or models.
   (c) Did you include any new assets in the supplemental material or as a URL? N/A.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? N/A.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. This information was conveyed to the human evaluators prior to the engagement.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? N/A
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? N/A
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes. Refer to "Appendix G - Evaluation Instructions"

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No. The human participants does not have prolonged exposure and are evaluating real-world memes objectively. Additionally, we did not collect any sensitive human data.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes. Refer to "Appendix H - Participant Compensation"

(d) Did you discuss how data is stored, shared, and deidentified? N/A.

Appendices

Appendix A - Limitations

Prompts Design. The experiments were conducted using a fixed set of prompts. This approach may raise questions about the sensitivity of the results to variations in prompts. However, the number of possible prompt variations is limitless, and analyzing the sensitivity of the results to different prompts would require additional human evaluations, which would increase both costs and time. The primary focus of this study is to demonstrate the utility of our framework on datasets with diverse themes, which has been effectively demonstrated. We leave the question of prompt optimality open, framing it as a prompt engineering challenge for future exploration.

Appendix B - Ethics Statement

Given that our work centers on meme corpora understanding rather than meme generation, it is unlikely to be exploited by malicious users. While extant public fears and sensitivities may be revealed by such analyses, if any, it is unlikely that they can be used by malicious users for targeted nefarious purposes. We do not anticipate significant negative societal impacts resulting from our work. On the contrary, it has the potential to provide decision-makers with a deeper and more nuanced understanding of public sentiment. This understanding can facilitate addressing public concerns and improving services. Nonetheless, there remains a need to improve the performance of the pipeline further before real-world deployment. If our pipeline fails to detect narratives of grievances, it risks overlooking critical concerns, which could undermine trust and result in dissatisfaction. Conversely, if the pipeline exaggerates or misrepresents concerns, it could result in the unnecessary allocation of resources by decision-makers to address non-existent issues.

Appendix C - Visual Representation of the MemeTT Pipeline

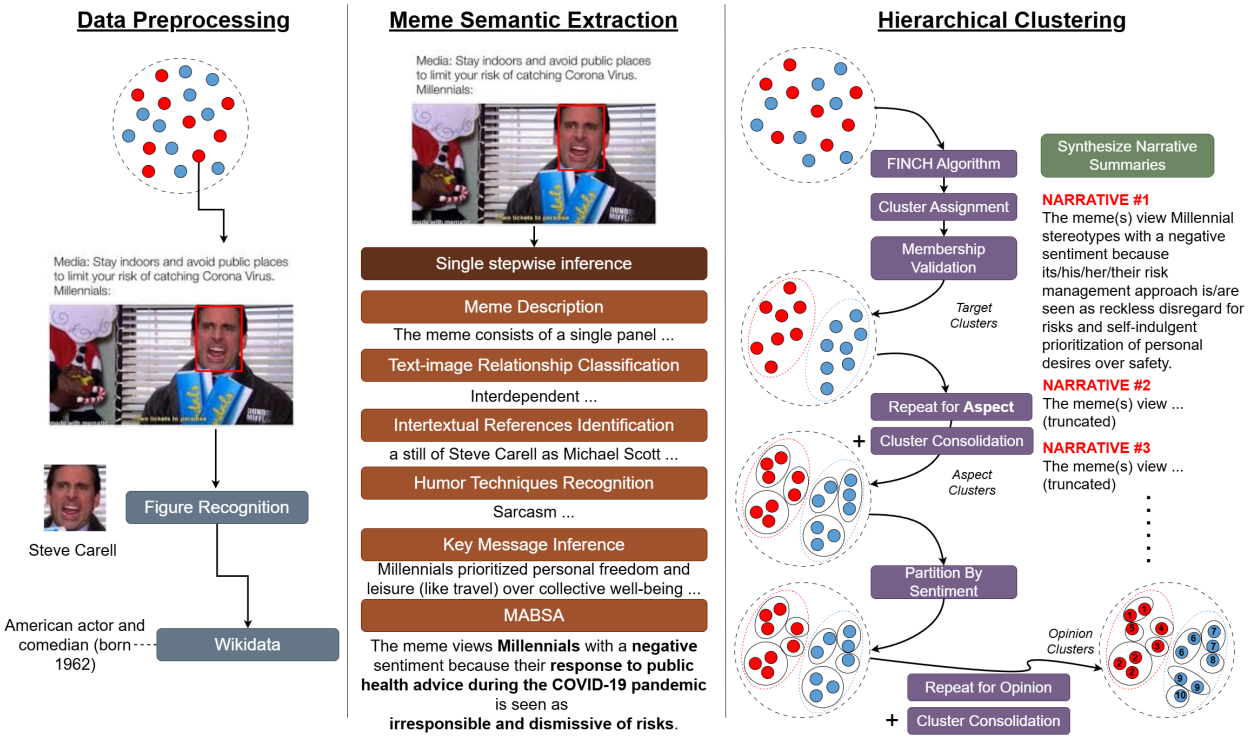


Figure 1: Visual representation of the MemeTT pipeline, illustrating the workflow from left to right.

Appendix D - Case Studies

	Memes	Quadruples	Human Eval.	LLM-Judge
(a)		**Target**: the practice of working from home **Aspect**: associated productivity or genuineness **Opinion**: questionable or potentially feigned **Sentiment**: negative	✓	✓
(b)		**Target**: misspelled internet slang as last words **Aspect**: use **Opinion**: absurdly trivializing **Sentiment**: negative	✗	✓
(c)		**Target**: the widespread aversion to tofu **Aspect**: extremity during panic buying **Opinion**: humorous **Sentiment**: neutral	✗	✗

Table 6: Case studies of agreement and disagreement comparing unanimous human evaluations versus unanimous LLM-Judge runs.

Observing that the LLM-Judge may be overly optimistic with respect to the accuracies of inferred semantic elements, we examined cases of disagreement. Our observations revealed that instances in which the LLM-Judge aligned with human evaluators often involved scenarios requiring general world knowledge rather than niche cultural understanding. Examples included widely relatable circumstances, such as the productivity of working from home during the COVID-19 pandemic (Table 6a) and panic-buying during COVID-19 (Table 6c). On the other hand, the LLM-Judge incorrectly assessed semantic elements as accurate when presented with memes featuring multi-layered jokes, specialized intertextual or inter-pictorial references, or allusions to non-Western culture. For example, it deemed a semantic quadruple correct (Table 6b) because it failed to understand the acronym "wgt ord lo" as "where got time Operationally Ready Date lo" (a catchphrase used by those who have completed their two-year service term to tease those who have not), instead interpreting it as misspelled internet slang.

Appendix E - Computational Cost and Settings

Model	Parameters	Input Tokens	Output Tokens	Input Cost	Output Cost
gemini-2.5-pro-exp-03-25	max_output_tokens=8192 seed=42 temperature=0 top_k=1 safety settings off	21.77M	24.11M	FREE	FREE
gemini-2.0-flash-thinking-exp-01-21	max_output_tokens=8192 seed=42 temperature=0 top_k=1 safety settings off	21.79M	8.93M	FREE	FREE
gemini-2.0-flash-001	max_output_tokens=8192 seed=42 temperature=0 top_k=1 safety settings off	21.76M	5.51M	0.15	0.60
llama4-maverick-instruct-basic	max_output_tokens=8192 temperature=0 top_k=1	27.63M	5.60M	0.22	0.88
llama4-scout-instruct-basic	max_output_tokens=8192 temperature=0 top_k=1	27.64M	6.37M	0.15	0.60
pixtral-large-2411	max_tokens=8192 temperature=0 random_seed=42	28.50M	8.67M	2.00	6.00
mistral-small-2503	max_tokens=8192 temperature=0 random_seed=42	20.00M	8.35M	0.10	0.30
pixtral-12b-2409	max_tokens=8192 temperature=0 random_seed=42	26.26M	6.73M	0.15	0.15
qwen2p5-vl-32b-instruct	max_output_tokens=8192 temperature=0 top_k=1	19.37M	12.06M	0.90	0.90
deepseek-r1-basic	max_output_tokens=32768 temperature=0 top_k=1	5.58M	2.47M	0.55	2.19
qwen3-235b-a22b	max_output_tokens=16384 temperature=0 top_k=1	4.53M	16.77M	0.22	0.88

Table 7: Model pricing and token usage. The token counts indicated for deepseek-r1-basic cover only the extraction of non-redundant viewpoints from outputs in the meme semantic extraction step for gemini-2.5-pro-exp-03-25. The token counts indicated for qwen3-235b-a22b cover only the clustering process using MABSA elements inferred by gemini-2.5-pro-exp-03-25.

Appendix F - Prompts

Process	Machine	t (hours)
Meme Semantic Extraction (InternVL2.5 8B MPO)	L40S	26.81
Meme Semantic Extraction (InternVL3 8B)	L40S	35.18
Meme Semantic Extraction (Qwen2.5 VL 7B)	L40S	44.22

Table 8: Computational costs of running experiments on Lightning AI Studio. Process = Step in the pipeline. Machine = Lightning AI machine used. t = time taken. As many steps in the pipeline are trivial in terms of time taken, we only include the ones that necessitated using GPUs.

MABSA (SYSTEM PROMPT)

[PURPOSE]:
System prompt to strictly enforce the constraints of Step 2a/4/6 when performing MABSA.

[PROMPT]:
```
Strictly and rigorously enforce, without exception, all the mandatory constraints of Step
2a (especially strictly enclosing the name of the image-text relationship type classified
in HTML <i></i> tags), Step 4 (especially strictly enclosing the name of each humor
technique present in the meme in HTML <b></b> tags), and Step 6 (especially strictly
stating viewpoints in the format of "The meme views {target noun or noun phrase} with a {
sentiment} sentiment because its/his/her/their {aspect noun or noun phrase} is/are seen as
 {opinion adjective or adjective phrase}." and with the sentiment being strictly either "
positive," or "neutral," or "negative."). Most importantly, firmly limit the number of
generated viewpoints to those that are truly accurate and central to the meme.
```

[REMARKS]:

MABSA (USER PROMPT PREFIX)

[PURPOSE]:
To provide identities (names) derived from Amazon Rekognition and associated descriptions of these identities derived from Wikidata as high-priority cues for person identification to facilitate MABSA.

[PROMPT]:
```
Pay extra close attention to the following:
"""
The person whose face is marked with a {color} bounding box (overlaid on the original meme
) is {name} ({description}).
"""
```

[REMARKS]:
- This prefix was used for the user prompt only when the Amazon Rekognition RecognizeCelebrities endpoint identified individuals in the meme with a MatchConfidence score of at least 90.
- The placeholders {color}, {name}, and {description} represented the color of the bounding box, the name of the identity derived from the Amazon Rekognition RecognizeCelebrities endpoint, and the corresponding Wikidata description, respectively.
- For each identified celebrity, we included exactly one sentence in the following format, with each sentence separated by a newline: "The person whose face is marked with a {color} bounding box (overlaid on the original meme) is {name} ({description})."

MABSA (USER PROMPT)

[PURPOSE]:
User prompt to implement a six-step workflow culminating in MABSA.

[PROMPT]:

Definitions of the types of image-text relationships whose names are enclosed in HTML tags and which are essential for Step 2a:
"""
Additive: where the text amplifies or elaborates on the image, or vice versa.
Synonymous: where the text and the image send essentially the same message.
Interdependent: where the image and the text together convey an idea that neither could convey alone.
Image dominant: where the image dominates, and the text does not add significantly to the meaning of the image.
Text dominant: where the image illustrates but does not add significantly to a largely complete text.
"""

Definitions of humor techniques whose names are enclosed in HTML tags and which are essential for Step 4:
"""
Comparison: Putting two or more elements together to produce a humorous situation.
Exaggeration: Overstating and magnifying something out of proportion.
Imitation: Mimicking or copying someone's appearance or movements.
Irony: Stating one thing and meaning something else or exactly the opposite of what is stated.
Parody: Imitating a style or a genre of literature or other media.
Personification: Attributing human characteristics to animals, plants, or objects.
Pun: Using elements of language to create new meanings that result in humor.
Sarcasm: Including blatant ironic responses or situations. Sarcasm features a more incisive tone than irony. Sarcasm also typically involves criticizing a target, while the same may not be true for irony.
Satire: Making a fool of or poking fun at well-known things, situations, or public figures.
Schadenfreude: Taking pleasure in other people's misfortune——victim humor.
Silliness: Making funny faces in ludicrous situations, showing silly or clownish behavior, or using silly voices as expressed in text, onomatopoeic sounds, or editing styles.
Surprise: Humor arises from unexpected situations.
"""

Definitions essential for identifying the viewpoints of the meme in Step 6:
"""
Target: A target may refer to one concrete, tangible entity or one abstract subject that was commented on in the meme. A target can be an individual, an organization, a community, a society, a government policy, a movement, a product, etc., and can be expressed as a named entity, a common noun, or a multi-word term.
Aspect: An aspect is one characteristic, attribute, or feature of the target of the meme.
Opinion: An opinion is an evaluation or attitude toward the aspect of the target of the meme.
Sentiment: The sentiment is the polarity of the opinion, either "positive," or "neutral," or "negative."
Viewpoint: A viewpoint comprises a target of the meme, an aspect of the target referenced, an opinion the meme expresses toward the aspect, and the sentiment polarity of the opinion. A meme may have one or more viewpoints because it may have one or more targets, with one or more aspects referenced for each target, and one or more opinions expressed toward each aspect.
"""

Step 1) Describe the meme.
a. If there is more than one panel, describe each panel in the meme individually.
b. Specify all objects and people in the meme, as well as backgrounds, scenery, interactions, and gestures or poses.
c. If there are multiple instances of any object or person, specify how many and where they are.
d. If people or characters are in the image, describe their facial expressions and the emotions they are conveying.
e. Mention what and where the text in the meme is and the font.

Step 2) Explain the function(s) of the text.

a. Classify and explain the type of image-text relationship the meme possesses. Each meme must only be classified into exactly one type.
b. If the text modifies your understanding of the image, explain exactly how and to what effect on the meaning of the meme.
Mandatory constraints for Step 2a:
"""
i. Strictly enclose the one name of the classified type of image-text relationship in HTML tags.
ii. Strictly avoid using Markdown syntax here.
"""

Step 3) Considering both the image and the text, identify and explain any specific references to:
a. popular culture events, artifacts and practices, including movie stills, video games, and other viral content;
b. political content, including photographs of political events and leaders, symbols, and statements by politicians;
c. national, cultural, social, and commercial symbols and icons;
d. country- or culture-specific jargon, acronyms and buzzwords; and
e. other memes.

Step 4) Identify all the humor technique(s) used in the meme. If there are any, explain how each humor technique is used and its effect on the meaning of the meme. If there are none, simply state so.
Mandatory constraints for Step 4:
"""
i. Strictly enclose the one-word name(s) of the humor technique(s) present in HTML tags.
ii. Strictly avoid using Markdown syntax here.
iii. Strictly do not mention humor techniques which are absent.
"""

Step 5) Considering all of the above and examining the meme as a whole, infer the meme's key intended message(s) within highly specific political, social or cultural contexts relevant to the meme.

Step 6) Extract and rewrite the meme's key intended message(s) (content from Step 5) as one or more highly specific and well-structured viewpoints.
Mandatory constraints for Step 6:
"""
i. A viewpoint must be expressed as one coherent and self-contained sentence faithfully following this strictly mandatory format:
"The meme views {target noun or noun phrase} with a {sentiment} sentiment because its/his/her/their {aspect noun or noun phrase} is/are seen as {opinion adjective or adjective phrase}."
ii. Every single placeholder——{target noun or noun phrase}, {aspect noun or noun phrase}, {opinion adjective or adjective phrase}, and {sentiment}——must be replaced with appropriately informative and specific word(s) to form a logical sentence.
iii. Opinion must be an informative and nuanced adjective (word) or short adjective phrase.
iv. Sentiment must strictly be classified as one of the three options, either "positive," or "neutral," or "negative."
v. Each viewpoint must comprise exactly one target, one aspect, one opinion, and one sentiment. Strictly do not conflate different entities as one target. Only if there are truly distinct viewpoints——for example, different targets, aspects, or opinions——generate separate viewpoints for each, ensuring they do not contradict one another. Otherwise, generate only one viewpoint. Firmly limit the number of generated viewpoints to those that are truly accurate and central to the meme.
vi. Strictly focus on what the meme actually conveys and nothing else.
"""

Strictly and rigorously follow all the instructions and the mandatory constraints above, including all the alphabetical instructions and all the mandatory constraints itemized using Roman numerals. Let's work this out in a step-by-step way to be sure we have perfectly accurate answers while rigorously adhering to the mandatory constraints for Step

2a, Step 4 and Step 6.

[REMARKS]:

- For the meme description, we refined the prompts employed by Singla et al. (2024) and Schuhmann and Bevan (2023).
- We adapted the taxonomy of text–image relationships proposed by McCloud (1994) and subsequently adopted by Yus (2019) for meme analysis. McCloud's taxonomy includes seven categories: *Word-specific*, *Picture-specific*, *Duo-specific*, *Additive*, *Parallel*, *Montage*, and *Interdependent*. We omitted the *Parallel* and the *Montage* categories, as Yus (2019) observed no such instances in his sample of memes. In practice, memes typically do not feature *Parallel* relationships, where the text and the image function independently without complementing each other, or *Montage*, where the text is embedded in the image instead of appearing as an overlay. Additionally, we updated the remaining terminology for greater clarity, for example, using the term *Synonymous* instead of *Duo-specific*.
- We adapted a question from Bettin et al. (2023) to encourage explicit reasoning about how the text contributes to the overall message.
- We included instructions in our prompt to elicit the identification of intertextual references prevalent in memes, such as cultural and political allusions (Mukhtar et al. 2024), movie stills and video game imagery (Lankshear and Knobel 2019), and photographs of public figures and events (Polách 2015).
- We drew on the humor type content analysis coding scheme proposed by Schumacher (2024), which in turn builds on the works of Catanescu and Tom (2001), Taecharungroj and Nueangjamnong (2015), and Buijzen and Valkenburg (2004).
- Our working definition of "target" in the MABSA context was synthesized from prior work (Alaei et al. 2023; Mohammad et al. 2016; Pontiki et al. 2015).
- Our working definition of "aspect" in the MABSA context was synthesized from prior work (Fuyao et al. 2023; Ayub et al. 2022).
- Our working definition of "opinion" in the MABSA context was synthesized from prior work (Huang et al. 2024).
- We extended the approach of Zhang et al. (2021), which formulates ABSA as a paraphrasing task, to MABSA.

EXTRACT VIEWPOINTS (SYSTEM PROMPT)

[PURPOSE]:

System prompt to extract viewpoints from MABSA outputs by VLMs.

[PROMPT]:

```
Instructions:
"""
Rigorously deduplicate (remove duplicates), extract, and standardize one or more non-redundant and explicitly stated viewpoint(s) from the meme analysis (Step 6) in the user message. Refer to what the meme analysis ultimately considers as the viewpoint(s) of the meme in the final analysis in Step 6. Rigorously follow, word for word, this mandatory format for each viewpoint:
"The meme views {target noun or noun phrase} with a {sentiment} sentiment because its/his/her/their {aspect noun or noun phrase} is/are seen as {opinion adjective or adjective phrase}."
Here, {target noun or noun phrase}, {sentiment}, {aspect noun or noun phrase} and {opinion adjective or adjective phrase} are placeholders that must be filled with meaningful content.
"""

Strict guidelines:
"""
1) When the provided viewpoint(s) do not already perfectly conform to the mandatory format, you may rewrite the viewpoint(s) by making cosmetic changes such as changes to the word form and the sentence structure while remaining highly faithful to the meaning of the original content, such that the extracted viewpoint(s) perfectly conform to the mandatory format. When the provided viewpoint(s) already perfectly conform to the mandatory format, you must strictly just copy and paste verbatim and be fully faithful to the explicitly stated viewpoint(s) in the meme analysis in the user message.
```

```
2) Rigorously enforce that {sentiment} is limited to only either "positive," or "neutral,"
 or "negative." If {sentiment} is not already either "positive," or "neutral," or "
negative," appropriately replace it strictly with only either "positive," or "neutral," or
 "negative."
3) Rigorously enforce that {aspect noun or noun phrase} is a specific noun word or noun
phrase and not just a simplistic pronoun.
4) A meme may have one or more viewpoints because it may have one or more targets, with
one or more aspects referenced for each target, and one or more opinions expressed toward
each aspect. However, any viewpoints that are mere close paraphrases of another viewpoint
must be removed and excluded from your answer. In particular, if there are multiple
provided viewpoints, compare the target-aspect-opinion-sentiment sets to determine if they
 are merely reworded variations that convey the same information. If multiple viewpoints
are extracted, they must be unique—having either different targets, aspects, or opinions—
and must not be mere close paraphrases of one another.
"""

Important answer format constraints:
"""
1) Strictly return a valid JSON with the key "viewpoint1" if there is only one
deduplicated viewpoint, or "viewpoint1," "viewpoint2," and so on if there is more than one
 deduplicated viewpoint.
2) Rigorously ensure, without exception, that {sentiment} is strictly either "positive,"
or "neutral," or "negative," and not any other words.
3) Keep your reasoning intelligent but concise.
4) Importantly, the placeholders {target noun or noun phrase}, {aspect noun or noun phrase
} and {opinion adjective or adjective phrase} must strictly be filled with meaningful
content and must not simply repeat the placeholder names themselves. That is, you must not
 use the generic terms "target(s)," "aspect(s)," or "opinion(s)" as the actual values for
these placeholders.
5) Most importantly, each viewpoint (JSON value) must rigorously follow, word for word,
the format of "The meme views {target noun or noun phrase} with a {sentiment} sentiment
because its/his/her/their {aspect noun or noun phrase} is/are seen as {opinion adjective
or adjective phrase}."
"""
```

[REMARKS]:

- We extended the approach of Zhang et al. (2021), which formulates ABSA as a paraphrasing task, to MABSA.

THEME SHORTLISTING FOR TARGET-LEVEL CLUSTERS (SYSTEM PROMPT)

[PURPOSE]:

System prompt to suggest a label for the cluster of meme targets.

[PROMPT]:

```
Propose one label that accurately and faithfully captures the highly-unique meaning of the
 cluster of semicolon-separated targets as defined in the user message. If it is possible
to propose a label that comfortably encapsulates all targets in the user message, try your
 very best to propose a label that encapsulates all targets. However, if trying to
encapsulate all targets will misrepresent the most frequent targets, strictly prioritize
representing the most frequent targets instead and ignore the peripheral targets. The
proposed label must not be a comma-separated list. The proposed label must be a noun or
noun phrase that is perfectly appropriate and grammatically fits as the value of the
placeholder {target noun or noun phrase} in the sentence: "The memes view {target noun or
noun phrase} with a {sentiment} sentiment because its/his/her/their {aspect noun or noun
phrase} is/are seen as {opinion adjective or adjective phrase}." Constraint: Think
intelligently and concisely. Strictly return only the label (noun or noun phrase) without
any preamble or anything extra.
```

[REMARKS]:

- We extended the approach of Zhang et al. (2021), which formulates ABSA as a paraphrasing task, to MABSA.

THEME SHORTLISTING FOR ASPECT-LEVEL CLUSTERS (SYSTEM PROMPT)

[PURPOSE]:
System prompt to suggest a label for the cluster of meme aspects.

[PROMPT]:
Propose one label that accurately and faithfully captures the highly-unique meaning of the cluster of semicolon-separated aspects of the target ({target}) as defined in the user message. If it is possible to propose a label that comfortably encapsulates all aspects in the user message, try your very best to propose a label that encapsulates all aspects. However, if trying to encapsulate all aspects will misrepresent the most frequent aspects, strictly prioritize representing the most frequent aspects instead and ignore the peripheral aspects. The proposed label must not be a comma-separated list. The proposed label must be a noun or noun phrase that is perfectly appropriate and grammatically fits as the value of the placeholder {{aspect noun or noun phrase}} in the sentence: "The memes view {target} with a {{sentiment}} sentiment because its/his/her/their {{aspect noun or noun phrase}} is/are seen as {{opinion adjective or adjective phrase}}." Constraint: Think intelligently and concisely. Strictly return only the label (noun or noun phrase) without any preamble or anything extra.

[REMARKS]:
• We extended the approach of Zhang et al. (2021), which formulates ABSA as a paraphrasing task, to MABSA.

THEME SHORTLISTING FOR OPINION-LEVEL CLUSTERS (SYSTEM PROMPT)

[PURPOSE]:
System prompt to suggest a label for the cluster of meme opinions.

[PROMPT]:
Propose one label that accurately and faithfully captures the highly-unique meaning of the cluster of semicolon-separated opinions toward the aspect ({aspect}) of the target ({target}) as defined in the user message. If it is possible to propose a label that comfortably encapsulates all opinions in the user message, try your very best to propose a label that encapsulates all opinions. However, if trying to encapsulate all opinions will misrepresent the most frequent opinions, strictly prioritize representing the most frequent opinions instead and ignore the peripheral opinions. The proposed label must not be a comma-separated list. The proposed label must be an adjective or adjective phrase that is perfectly appropriate and grammatically fits as the value of the placeholder {{opinion adjective or adjective phrase}} in the sentence: "The memes view {target} with a {sentiment} sentiment because its/his/her/their {aspect} is/are seen as {{opinion adjective or adjective phrase}}." Constraint: Think intelligently and concisely. Strictly return only the label (adjective or adjective phrase) without any preamble or anything extra.

[REMARKS]:
• We extended the approach of Zhang et al. (2021), which formulates ABSA as a paraphrasing task, to MABSA.

MEMBERSHIP VALIDATION (SYSTEM PROMPT)

[PURPOSE]:
System prompt to detect labels that are completely irrelevant to the original semantic elements.

[PROMPT]:
Think intelligently and concisely. Strictly return only one single word, either "Yes" or "No" without any explanation or anything extra.

[REMARKS]:

MEMBERSHIP VALIDATION (USER PROMPT)

[PURPOSE]:
User prompt to detect labels that are completely irrelevant to the original semantic elements.

[PROMPT]:
```
Is "{text}" absolutely irrelevant and unrelated to the theme of "{label}"?
```

[REMARKS]:

THEME CONSOLIDATION (USER PROMPT)

[PURPOSE]:
User prompt to group themes.

[PROMPT]:
```
In the user message is a semicolon-separated list of cluster labels. Your task is to
identify any provided labels that are either (i) absolutely synonymous with one or more
other labels (i.e., they are, without question, identical in meaning); or (ii) clearly
encompass one or more other labels in terms of meaning. Strictly output a valid JSON
object with each key being one synonymous/encompassing label, and each value being a
semicolon-separated list of the corresponding synonymous/encompassed label(s). For example
, given "anger; happy; sad; sorrowful; fear; elated; joyful; dejected; emotions; joy", the
 output should be {"happy": "elated; joyful", "sad": "sorrowful; dejected", "emotions": "
anger; fear; joy"}. However, if no synonymous/encompassing labels are found, simply output
 {"None": "None"}. Constraint: Think intelligently and concisely. Strictly enforce that
each key is either (i) identical in meaning and nuance with the label(s) in its value, or
(ii) very clearly encompasses the meaning of the label(s) in its value. Strictly enforce
and double-check that each provided label in the user message must not appear more than
once in your JSON answer.
```

[REMARKS]:

LLM-JUDGE OF MABSA ACCURACY (SYSTEM PROMPT)

[PURPOSE]:
System prompt to assess MABSA accuracy.

[PROMPT]:
```
Definitions:
"""
Target: A target may refer to one concrete, tangible entity or one abstract subject that
was commented on in the meme. A target can be an individual, an organization, a community,
 a society, a government policy, a movement, a product, etc., and can be expressed as a
named entity, a common noun, or a multi-word term.
Aspect: An aspect is one characteristic, attribute, or feature of the target of the meme.
Opinion: An opinion is an evaluation or attitude toward the aspect of the target of the
meme.
Sentiment: The sentiment is the polarity of the opinion: "positive," or "neutral," or "
negative."
Viewpoint: A viewpoint comprises a target of the meme, an aspect of the target referenced,
 an opinion the meme expresses toward the aspect, and the sentiment polarity of the
opinion.
"""

Instructions: Evaluate the quadruple of semantic elements (target, aspect, opinion,
sentiment) by first assessing the accuracy of the target with respect to the meme. If the
target is accurate, assess the aspect; if the aspect is accurate, assess the opinion; and
if the opinion is accurate, assess the sentiment. If at any point a semantic element is
```

not accurate, there is no need to proceed further with the quadruple. A semantic element is accurate if it (i) is mentioned in the meme, whether implicitly or explicitly; (ii) does not distort the message of the meme; (iii) effectively conveys the intended message of the meme.

Constraint: Answer strictly with a single number: 0 (no elements accurate) or 1 (only target accurate) or 2 (only target and aspect accurate) or 3 (only target, aspect, opinion accurate) or 4 (all elements accurate). Return the number without anything else.

[REMARKS]:

- Our working definition of "target" in the MABSA context was synthesized from prior work (Alaei et al. 2023; Mohammad et al. 2016; Pontiki et al. 2015).
- Our working definition of "aspect" in the MABSA context was synthesized from prior work (Fuyao et al. 2023; Ayub et al. 2022).
- Our working definition of "opinion" in the MABSA context was synthesized from prior work (Huang et al. 2024).

LLM-JUDGE OF CLUSTER HOMOGENEITY (SYSTEM PROMPT)

[PURPOSE]:

System prompt to assess cluster homogeneity.

[PROMPT]:

```
Definitions:
"""
Target: A target may refer to one concrete, tangible entity or one abstract subject that
was commented on in the meme. A target can be an individual, an organization, a community,
 a society, a government policy, a movement, a product, etc., and can be expressed as a
named entity, a common noun, or a multi-word term.
Aspect: An aspect is one characteristic, attribute, or feature of the target of the meme.
Opinion: An opinion is an evaluation or attitude toward the aspect of the target of the
meme.
Sentiment: The sentiment is the polarity of the opinion: "positive," or "neutral," or "
negative."
Viewpoint: A viewpoint comprises a target of the meme, an aspect of the target referenced,
 an opinion the meme expresses toward the aspect, and the sentiment polarity of the
opinion. A viewpoint can be expressed as one coherent and self-contained sentence
following this format:
"The meme views {{target noun or noun phrase}} with a {{sentiment}} sentiment because its/
his/her/their {{aspect noun or noun phrase}} is/are seen as {{opinion adjective or
adjective phrase}}."
"""

Instructions: You have been given exactly {n} meme images in the user message. Using the
definitions provided above, determine the viewpoint(s) of each meme. A meme may have one
or more viewpoints because it may have one or more targets, with one or more aspects
referenced for each target, and one or more opinions expressed toward each aspect. Your
goal is to find the largest group of memes, from the {n} provided, that share the exact
same viewpoint. For viewpoints to be identical, the target of the meme, the aspect of the
target referenced, the opinion expressed about that aspect, and the sentiment polarity of
the opinion must all be exactly the same. If even one of the elements (targets or aspects
or opinions or sentiments) is dissimilar between the memes, they must not be grouped
together; their viewpoints must be deemed different. Find the one specific viewpoint that
appears in the highest number of memes. Then, count how many memes share that top
viewpoint.

Constraint: Return a single integer (number of memes that share the top viewpoint) and
nothing else. Your answer must be an integer between 1 and {n}, inclusive. If no viewpoint
 is shared by more than one meme, return 1.
```

[REMARKS]:

- Our working definition of "target" in the MABSA context was synthesized from prior work

(Alaei et al. 2023; Mohammad et al. 2016; Pontiki et al. 2015).
- Our working definition of "aspect" in the MABSA context was synthesized from prior work (Fuyao et al. 2023; Ayub et al. 2022).
- Our working definition of "opinion" in the MABSA context was synthesized from prior work (Huang et al. 2024).

LLM-JUDGE OF INTER-CLUSTER QUALITY (SYSTEM PROMPT)

[PURPOSE]:
System prompt to assess inter-cluster quality.

[PROMPT]:
```
Definitions:
"""
Target: A target may refer to one concrete, tangible entity or one abstract subject that
was commented on in the meme. A target can be an individual, an organization, a community,
 a society, a government policy, a movement, a product, etc., and can be expressed as a
named entity, a common noun, or a multi-word term.
Aspect: An aspect is one characteristic, attribute, or feature of the target of the meme.
Opinion: An opinion is an evaluation or attitude toward the aspect of the target of the
meme.
Sentiment: The sentiment is the polarity of the opinion: "positive," or "neutral," or "
negative."
"""

Instructions: Which of the following memes LEAST FITS the narrative in the user message?
Select only one. Fit is your subjective assessment of whether the meme's target, aspect,
opinion, and sentiment are appropriately represented by the narrative, and not simply
unrelated to the narrative. Examine the entire narrative and compare it with each meme
before making the selection.

Constraint: Answer strictly with a single alphabet corresponding to the meme: "A" or "B"
or "C."
```

[REMARKS]:
- Our working definition of "target" in the MABSA context was synthesized from prior work (Alaei et al. 2023; Mohammad et al. 2016; Pontiki et al. 2015).
- Our working definition of "aspect" in the MABSA context was synthesized from prior work (Fuyao et al. 2023; Ayub et al. 2022).
- Our working definition of "opinion" in the MABSA context was synthesized from prior work (Huang et al. 2024).

LLM-JUDGE OF INTRA-CLUSTER QUALITY (COHERENCE) (SYSTEM PROMPT)

[PURPOSE]:
System prompt to assess intra-cluster quality (coherence).

[PROMPT]:
```
Rate the COHERENCE of the narrative in the user message based on how well the target, the
aspect, and the opinion in the narrative fit together. When judging coherence, ignore the
possessive determiners and subject-verb agreement; regard "is/are" and "its/his/her/their"
 as template artifacts and do not penalize coherence or fluency for them.
1. Incoherent: The target, the aspect, and the opinion are all completely unrelated and do
 not fit together at all.
2. Mostly Incoherent: Only two of the three elements (target, aspect, opinion) are related
 and fit together.
3. Mostly Coherent: All three elements (target, aspect, opinion) are related and fit
together, but the phrasing is awkward.
4. Highly Coherent: All three elements (target, aspect, opinion) are related and fit
together, and the narrative is fluent.
```

Constraint: Answer strictly as a number without anything extra, 1 or 2 or 3 or 4.
[REMARKS]:

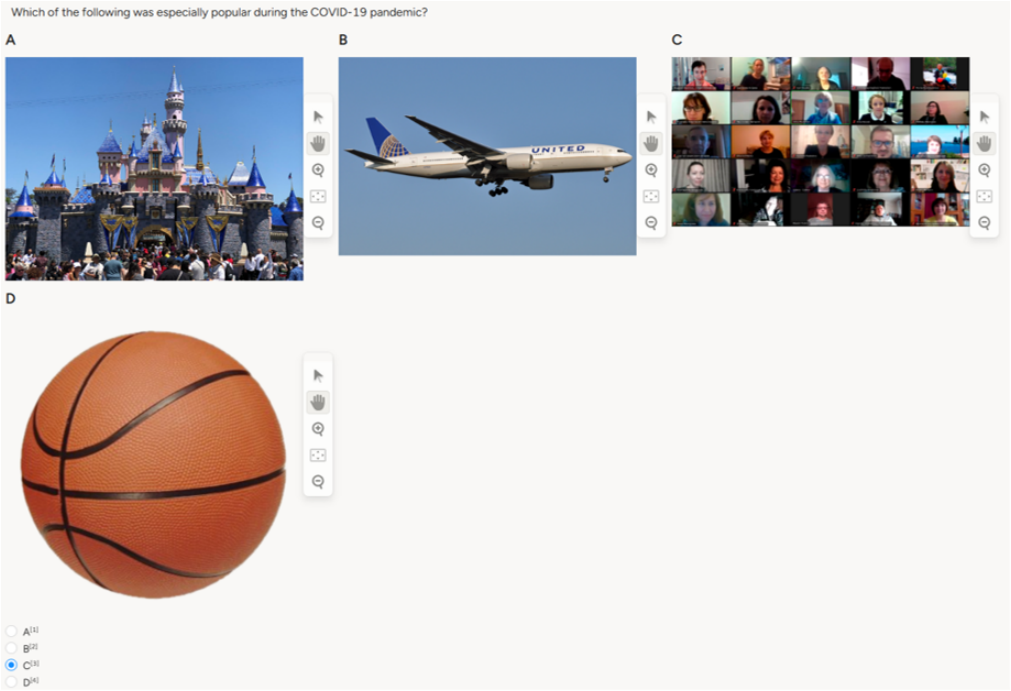
LLM-JUDGE OF INTRA-CLUSTER QUALITY (RELEVANCE) (SYSTEM PROMPT)

[PURPOSE]:
System prompt to assess intra-cluster quality (relevance).

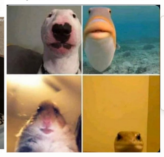
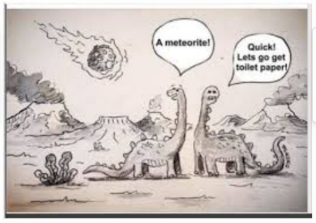
[PROMPT]:
Thoroughly read the meme (the image passed in the user message). Carefully read the narrative (comprising the target, the aspect, the opinion, and the sentiment) in the user message (text prompt) and compare it with the meme (the image passed in the user message).
Q1. Identify whether (true or false) the narrative is relevant to the target. Here, relevance refers to whether the narrative in the user message captures the target of the meme (the image in the user message).
Q2. If Q1 is true, identify whether (true or false) the narrative is relevant to the aspect. Here, relevance refers to whether the narrative in the user message captures the aspect of the target of the meme (the image in the user message). Slight generalization of the aspect is permitted, as long as it remains accurate and informative. If Q1 is false, set Q2 to false.
Q3. If Q2 is true, identify whether (true or false) the narrative is relevant to the opinion and sentiment. Here, relevance refers to whether the narrative in the user message captures the opinion and sentiment toward the aspect of the target of the meme (the image in the user message). If Q2 is false, set Q3 to false.
Constraint: Answer strictly as a valid JSON object, with three keys: "Q1" and "Q2" and "Q3." The values of "Q1" and "Q2" and "Q3" must be strictly BOOLEAN (true or false) only.

[REMARKS]:

Appendix G - Evaluation Instructions

Purpose Of Instruction	Example Screenshot
Evaluate basic knowledge about important themes in the meme corpora	

Continued on next page.

Purpose Of Instruction	Example Screenshot
Evaluate accuracy of semantic elements from MABSA	 Click_here_for_meme Evaluate the accuracy of each semantic element in each quadruple (target, aspect, opinion, sentiment). A semantic element is accurate if it (i) is mentioned in the meme, whether implicitly or explicitly; (ii) does not distort the message of the meme; (iii) effectively conveys the intended message of the meme. Viewpoint: The meme views people accidentally revealing their unprepared appearance on webcam during conference calls with a negative sentiment because their appearance is seen as potentially embarrassing and comically unpresentable. Target: people accidentally revealing their unprepared appearance on webcam during conference calls Aspect: appearance Opinion: potentially embarrassing and comically unpresentable Sentiment: negative target[1] aspect[2] opinion[3] sentiment[4]
Evaluate inter-cluster quality	Which of the following memes LEAST FITS the narrative? Select only one. Fit is your subjective assessment of whether the meme's target, aspect, opinion and sentiment are appropriately represented by the narrative, and not simply unrelated to the narrative. Narrative: The meme(s) view panic buying behavior during crises with a negative sentiment because its/his/her/their irrational hoarding of non-essential items is/are seen as absurd and irrational. Examine the entire narrative and compare it with each meme before making the selection. Click here for meme A Click here for meme B Click here for meme C A   B  C  A[1] B[2] C[3]
Evaluate coherence of narrative	Narrative: The meme(s) view handwashing with a positive sentiment because its/his/her/their fundamental hygiene requirement is/are seen as hyperbolically necessary. Target: handwashing Aspect: fundamental hygiene requirement Opinion: hyperbolically necessary Sentiment: positive Q1. Rate the COHERENCE of the narrative based on how well the target, aspect and opinion in the narrative fit together. 1. Incoherent: The target, the aspect and the opinion are all completely unrelated and do not fit together at all.[1] 2. Mostly Incoherent: Only two of the three elements (target, aspect, opinion) are related and fit together.[2] 3. Mostly Coherent: All three elements (target, aspect, opinion) are related and fit together, but the phrasing is awkward.[3] 4. Highly Coherent: All three elements (target, aspect, opinion) are related and fit together and the narrative is fluent.[4]

Continued on next page.

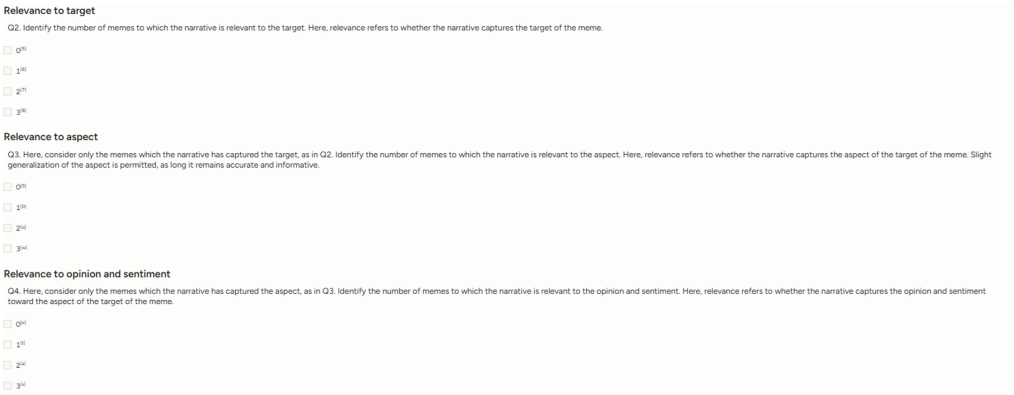
Purpose Of Instruction	Example Screenshot
Encourage a meticulous examination of memes before assessing relevance	
Evaluate relevance of narrative to memes	

Table 9: Purpose of instructions given to participants and example screenshots of instructions.

Appendix H - Participant Compensation

We paid participants a rate of S$20/hour (approximately US$15.60/hour). On average, each participant took 12.5 hours to complete the tasks.

Appendix I - Case Normalization Issue

Case normalization was not applied consistently during downstream post-processing of cluster labels. As a result, labels differing only in capitalization (e.g., "the Democratic Party" vs. "The Democratic Party") were sometimes treated as distinct groups in later clustering steps. In a post hoc audit of the *Gemini 2.5 Pro* MABSA clustering output, we identified that this issue affected the clustering of 693 of 12,242 quadruples/viewpoints (5.66%) across 264 narratives. Importantly, this issue is confined to label formatting in downstream grouping steps and does not affect semantic extraction. We also audited all clustering-related human evaluation samples and found that 34 of 723 samples (4.70%) involved narratives affected by the capitalization issue in the main clustering output. While this identifies potentially affected samples, we did not estimate how many human judgments would actually change under a corrected rerun. To verify that this issue does not materially affect our findings, we recalculated all evaluation metrics after excluding the affected narratives. The results, reported in Tables 10, 11, 12, and 13, are consistent with those in the main paper. These audits indicate a limited post-processing issue with label capitalization, with the overall conclusions remaining unchanged.

Clustering	Num. C	Macro-Purity	Micro-Purity
MemeTT			
Gemini 2.5 Pro (F)†	5546	.830	.753

Table 10: Cluster homogeneity evaluation after excluding narratives affected by the case-normalization issue (mean of three LLM-Judge runs). Num. C = the number of clusters. The dagger sign (†) indicates that MemeTT yields singleton clusters, each contributing purity 1.

		Coherence					
$\|C\|$	N	Hum.	LLM-J	Agr_h	Agr_{hl}	$AC1_h$	$AC1_{hl}$
Harm-C							
3	47	3.70	3.92	23	23	.591	.652
4	50	3.71	3.86	18	17	.524	.596
5	22	3.77	3.90	10	10	.622	.714
Harm-P							
3	45	3.91	3.94	35	34	.836	.864
4	42	3.94	3.97	35	35	.884	.910
5	36	3.94	3.99	32	32	.924	.938
TDMeme							
3	49	3.62	3.97	15	15	.428	.536
4	39	3.66	3.93	14	13	.500	.577
5	15	3.67	3.91	5	5	.530	.566

Table 11: Cluster narrative coherence evaluation after excluding narratives affected by the case-normalization issue (LLM-Judge values are means of 5 runs). $|C|$ = Number of memes in cluster. N = Number of clusters assessed. Hum. = Average human score. LLM-J = GPT-5 mini (average score). Agr_h = The number of instances in which all human annotators independently assigned the maximum score of 4. Agr_{hl} = The number of instances in which all human annotators and the LLM-Judge (majority vote of 5 runs) independently assigned the maximum score of 4. $AC1_h$ = Gwet's AC1 between human evaluators. $AC1_{hl}$ = Gwet's AC1 treating the majority vote of the LLM-Judge as an additional evaluator.

		Target				Aspect				Opinion-Sentiment			
$\|C\|$	N	Hum.	LLM-J	Agr_h	Agr_{hl}	Hum.	LLM-J	Agr_h	Agr_{hl}	Hum.	LLM-J	Agr_h	Agr_{hl}
Harm-C													
3	47	2.80	2.84	31	29	2.52	2.69	18	15	2.28	2.20	10	6
4	50	3.64	3.81	29	28	3.39	3.64	21	19	3.09	3.10	16	11
5	22	4.48	4.75	12	12	4.05	4.53	6	6	3.58	3.85	3	2
Harm-P													
3	45	2.82	2.85	36	36	2.45	2.44	19	15	2.38	2.03	19	12
4	42	3.79	3.71	30	27	3.56	3.15	18	12	3.41	2.61	15	6
5	36	4.76	4.74	20	20	4.39	4.32	11	10	4.13	3.79	10	9
TDMeme													
3	49	2.62	2.68	21	19	2.10	2.16	5	4	1.80	1.83	2	2
4	39	3.62	3.67	21	21	2.89	3.18	3	3	2.34	2.61	0	0
5	15	4.42	4.60	7	6	3.33	3.57	0	0	2.71	2.51	0	0

Table 12: Relevance of cluster narratives to the quadruple elements reflected in each cluster's memes after excluding narratives affected by the case-normalization issue (LLM-Judge values are means of 5 runs). $|C|$ = Number of memes in cluster. N = Number of clusters assessed. Hum. = Average human score. LLM-J = GPT-5 mini (average score). Agr_h = The number of instances in which all human annotators independently assigned the maximum score of $|C|$, Agr_{hl} = The number of instances in which all human annotators and the LLM-Judge (majority vote of 5 runs) independently assigned the maximum score of $|C|$.

Dataset	N	Trg.	Asp.	Snt.	Human	LLM-Judge	α_h	α_{hl}
Harm-C	30	✗	✗	✗	96.67%	96.67%	.831	.855
	29	✓	✗	✗	68.97%	72.41%	.606	.557
	29	✓	✓	✗	62.07%	65.52%	.339	.356
	30	✓	✓	✓	43.33%	30.00%	.345	.322
Harm-P	29	✗	✗	✗	93.10%	86.21%	.966	.923
	28	✓	✗	✗	71.43%	85.71%	.492	.575
	29	✓	✓	✗	55.17%	82.76%	.365	.372
	23	✓	✓	✓	47.83%	65.22%	.391	.391
TDMeme	30	✗	✗	✗	93.33%	96.67%	.718	.750
	30	✓	✗	✗	50.00%	63.33%	.298	.392
	27	✓	✓	✗	37.04%	51.85%	.447	.424
	30	✓	✓	✓	43.33%	70.00%	.517	.443

Table 13: Inter-cluster quality experiment results after excluding narratives affected by the case-normalization issue. N = Number of samples assessed. Trg. = Target. Asp. = Aspect. Snt. = Sentiment. Human = Human (majority vote). LLM-Judge = GPT-5 mini (majority vote). α_h = Krippendorff's Alpha between human evaluators. α_{hl} = Krippendorff's Alpha treating the majority vote of the LLM-Judge as an additional evaluator.