MemeMatch: A Large-Scale Dual-Context Multimodal Dataset and Retrieval System for Internet Memes

Do Tri An Le[1], Donát Ákos Köller[2], Qixin Deng[1*], Roland Molontay[2*]

[1]Department of Mathematics and Computer Science, Wabash College, Indiana, USA
[2]Department of Stochastics, Budapest University of Technology and Economics, Budapest, Hungary
tdle26@wabash.edu, kollerd@edu.bme.hu, dengq@wabash.edu, molontay@math.bme.hu

Abstract

We introduce **MemeMatch**, a large-scale multimodal meme dataset and retrieval system that bridges meme collection, annotation, and analysis in a unified pipeline. The dataset contains nearly one million image-with-text memes from Reddit's `r/Memes` (2018–2023) and ImgFlip, with rich metadata. Each meme is decomposed into two semantic contexts: *local context*, capturing the editable text payload (overlay text and title), and *global context*, capturing the underlying visual substrate or template semantics. Both are enriched with transformer-based annotations, including 14-dimensional sentiment and emotion vectors, BERTopic-derived topics, and zero-shot usage-intent labels. This structured representation supports exploratory analysis and context-aware retrieval by natural language or image query.

Introduction

Memes are a cornerstone of online communication, encapsulating humor, commentary, and emotion in compact multimodal form (Wang et al. 2025). Their rapid diffusion across platforms makes them key artifacts for understanding digital culture and public sentiment. However, their analysis remains challenging because meme meaning arises from an interplay between image and text, often relying on cultural references, irony, and shared social context, as highlighted by recent studies on topicality and multimodal content dynamics in social media (Barnes et al. 2024).

Previous computational work on image-with-text memes has often been shaped by available benchmark datasets. The Facebook Hateful Memes dataset (Kiela et al. 2020) and the SemEval Memotion benchmark (Sharma, Bhageria, and Das 2020) introduced multimodal classification tasks on a few thousand labeled memes, focusing on sentiment, humor, or hate detection. While valuable, these datasets are limited in scale and diversity, often lacking fine-grained emotional or contextual annotations. Other studies have explored CLIP-based multimodal embeddings for meme clustering or retrieval (Radford et al. 2021), and recent approaches for meme classification underscore the benefits of fusing text and image signals (Huertas-Tato et al. 2024). These efforts confirm the need to integrate modalities, yet a large-scale, richly annotated dataset capturing real-world meme diversity has remained absent.

MemeMatch addresses this gap by combining Reddit and ImgFlip into a large-scale multimodal meme corpus. Reddit provides organically shared memes from online communities, while ImgFlip provides standardized templates and user-created variations. Our main contributions are:

- A **large-scale meme dataset** ($\sim$301K memes across 2,083 templates) combining organic social memes and template-based examples from two major platforms.
- A **dual-context representation** separating *local context* (user overlay + title) from *global context* (template caption), preserving both message and form (Dubey et al. 2018).
- Rich **automatic annotations** including sentiment/emotion vectors (14 dimensions), BERTopic clusters (300 topics for the local context and 200 for the global context), and 28 usage intent labels.
- A **context- and intent-aware retrieval system** supporting natural language and image queries via a framework of precomputed case-based embeddings and an LLM-based query parser.
- Exploratory findings on **meme affect, topics, and usage patterns**.

This work provides both a reusable research resource and an applied system. Further details are available in our GitHub repository[1]. By unifying scalable meme annotation, representation, and retrieval, MemeMatch enables robust analyses of how humor, affect, and culture interact across social media.

Methodology

MEMEMATCH follows an end-to-end multimodal workflow that connects acquisition, preprocessing, and analytic enrichment into a unified data–system pipeline. Memes are collected from Reddit and Imgflip with metadata (e.g., timestamp, upvotes). After duplicate removal and normalization, each meme is decomposed into two semantic layers: **local context** (user-added overlay text) and **global context** (the underlying visual template). Both layers are enriched with emotion, sentiment, topic, and usage annotations using

*Corresponding authors.

[1]https://github.com/TriAnLe171/MemeMatch-v1.0

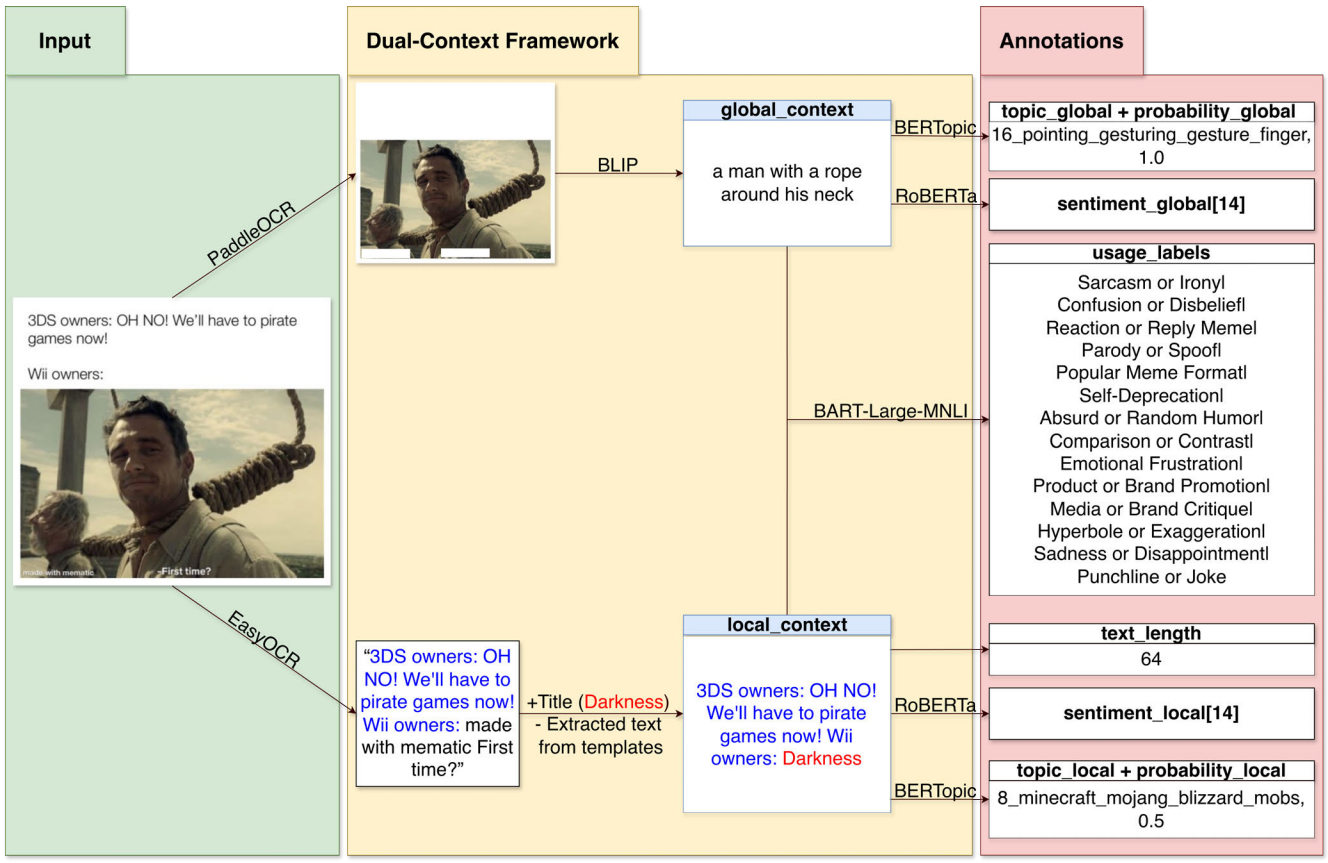


Figure 1: End-to-end Dual-Context pipeline: EasyOCR + Reddit title yield the local text (`local_context`), while PaddleOCR masking with BLIP yields a template-grounded global caption (`global_context`); both contexts are annotated for topics (BERTopic), affect (RoBERTa), usage intent (BART-MNLI), and auxiliary features.

transformer-based models. The resulting dataset supports two downstream paths: (1) embedding generation for similarity search and context-aware retrieval, and (2) exploratory data analysis of affective and cultural trends.

Data Collection and Preprocessing

Our data comes from two sources: ImgFlip and Reddit.

ImgFlip is a popular website for creating and sharing image-with-text and GIF memes, offering many customizable templates (preexisting visual layouts). Because each meme is explicitly labeled with its template, ImgFlip is well suited for supervised learning. In contrast, Know Your Meme (KYM) lacks reliable template-to-instance mappings, as its example images are often thematically related rather than direct template outputs. Using Beautiful Soup (Richardson 2013), we collected 3,127 ImgFlip templates (April 2025); after deduplication, 2,083 unique templates remained. Scraping each template library yielded 153,792 memes labeled by template, averaging 74 memes per template.

Reddit, often described as *the front page of the internet* (Sanderson and Rigby 2013), is a major social media platform and rich source of viral memes. We collected posts from `r/Memes` (over 26M subscribers) from Jan. 2018 to Dec. 2023 with metadata (title, upload date, upvotes). Upvote counts reflect the crawl window (Jan. 2024) rather than lifetime totals. Using the Pushshift Reddit API (Baumgartner et al. 2020) with PRAW and PMAW (Boe 2024; Ludwig 2024), we sampled up to 1K random posts per day, yielding 899,522 images.

To improve dataset integrity, duplicates were identified using perceptual hashing (pHash, (Klinger and Starkweather 2008)) and near-identical images were merged, keeping one representative per group. After preprocessing, we retain **146,991** unique Reddit memes without pre-labeled templates and **153,792** ImgFlip memes across **2,083** unique templates; this combined corpus is the input to our Dual-Context Framework.

Dual-Context Framework

Each meme is processed through two synchronized branches, yielding two complementary semantic contexts (Figure 1). The **local context** captures user-generated text and situational meaning, while the **global context** captures template-level visual semantics.

Local Context Extraction EasyOCR (Jaided AI 2020) extracts overlay text from meme images. To remove template

artifacts, we OCR 2,083 common Imgflip templates and filter overlapping strings, eliminating repeated or non-informative tokens (e.g., watermarks). The cleaned text is concatenated with the Reddit title from the core metadata to form the final `local_text` field (see Figure 1, Dual-Context Framework).

Global Context Extraction User-added text regions are localized with PaddleOCR (Du et al. 2020) and masked prior to captioning with BLIP (Li et al. 2022). BLIP then generates a template-grounded description of the visual content, stored as `global_context` (see Figure 1, Dual-Context Framework). Notably, Reddit memes are organic and not necessarily drawn from canonical meme templates; here, we use "template" to refer broadly to the underlying base image after text masking.

Implementation note. We adopt a hybrid OCR setup because extraction and localization have different requirements: EasyOCR provides fast, robust text extraction at scale, while PaddleOCR yields tighter bounding boxes for masking overlay text prior to BLIP captioning.

A qualitative inspection of 50 random memes spanning 10 templates suggests that memes sharing the same template often have similar global context but different local context, affect, and usage labels due to changes in overlay text. This separation helps distinguish reusable visual form from post-specific meaning.

Output The Dual-Context Framework produces per-meme `local_context` and `global_context` fields that feed the subsequent Annotations stage (Figure 1).

Annotations

Building on the last stage, the Annotations stage transforms each meme's textual `local_context` and `global_context` into standardized semantic signals.

Sentiment Modeling To capture affective tone, we use two RoBERTa-based models fine-tuned on Twitter data (Cardiff NLP (Camacho-collados et al. 2022; Loureiro et al. 2022)):

- `twitter-roberta-base-emotion-multilabel-latest`: predicts 11 emotion categories.
- `twitter-roberta-base-sentiment-latest`: classifies *positive*, *neutral*, and *negative* polarity.

Each meme yields a 14-dimensional affect vector (11 emotions + 3 polarities) for both local and global contexts, capturing fine-grained emotions (e.g., *joy*, *anger*) and coarse sentiment polarity. As shown in Table 1, the model outputs class-wise confidence scores for a local-context example. We retain the full Negative–Neutral–Positive confidence vector rather than collapsing it to a scalar, since memes often convey ambiguity, mixed affect, or irony better captured by class-wise probabilities.

Usage Labeling To infer communicative intent, we applied zero-shot classification using `facebook/bart-large-mnli` across 28 predefined usage labels (e.g., *Sarcasm or Irony*, *Confusion or Disbelief*, *Parody or Spoof*). Local and global texts are concatenated; per meme, labels with model confidence ≥ 0.70 are retained. Figure 1 illustrates an example prediction produced by the Annotations stage.

Emotion \ Text	"3DS owners: OH NO! We'll have to pirate games now!" Wii owners:"
Anger	0.479241
Anticipation	0.199450
Disgust	0.647908
Fear	0.066868
Joy	0.032594
Love	0.002387
Optimism	0.016223
Pessimism	0.075464
Sadness	0.217513
Surprise	0.060239
Trust	0.005590
Negative	0.768447
Neutral	0.215292
Positive	0.016261

Table 1: Per-label confidence scores (11 emotions + 3 sentiment polarities) for the `local_context` of the meme sample shown in Figure 1; higher values indicate stronger model confidence.

Topic Modeling We use BERTopic (Grootendorst 2022) to model themes in local and global text. Texts are embedded with `all-mpnet-base-v2` (Sentence-Transformers), reduced via UMAP (`n_neighbors=20`, `n_components=5`, `metric='cosine'`) (McInnes et al. 2018), and clustered with HDBSCAN (`min_cluster_size=30`) (McInnes, Healy, and Astels 2017). We then extract keywords using a KeyBERT–MMR hybrid (Grootendorst 2020) to produce concise, less redundant topic summaries. We set `nr_topics=300` for local context and `nr_topics=200` for global context. About 87.7% of memes fall into high-confidence clusters; the rest are labeled as outliers (`-1`). An example is shown in Figure 1 (Annotations stage).

Auxiliary Features We compute `text_length` (character count in `local_context`) as a lightweight auxiliary feature that complements higher-level annotations and supports downstream analysis.

Output The Annotations stage yields a dual-context annotated dataset: for each meme we store local/-global affect vectors (`sentiment_local[14]`, `sentiment_global[14]`), topic assignments with confidences (`topic_local`, `probability_local`; `topic_global`, `probability_global`), usage labels (`usage_labels`), and auxiliary fields (e.g., `text_length`). These annotations underpin downstream retrieval and exploratory analyses.

Data Structure

This section formalizes the released artifact for MEMEMATCH v1.0. The schema mirrors the Methodology:

core metadata from crawl time and derived annotations from the Dual-Context Framework and Annotations stage (Figure 1). Tables 2 and 3 list all fields, followed by data-quality notes and parsing guidelines.

Schema

Field	Type	Req.	Description
filename	string	✓	Image filename, used as the unique identifier across all tables. For ImgFlip memes, the filename also indicates the template name.
created_utc	string		Reddit meme upload timestamp (ISO-8601 format).
score	int		Reddit upvote count at crawl time.

Table 2: Schema (core metadata) for MEMEMATCH v1.0.

Field	Type	Req.	Description
local_context	string	✓	Cleaned OCR overlay text concatenated with title.
global_context	string	✓	BLIP caption on masked image (template semantics).
text_length	int	✓	Number of characters in `local_text`.
sentiment_local[14]	float[0..1]	✓	11 emotions + 3 polarities for local text.
sentiment_global[14]	float[0..1]	✓	11 emotions + 3 polarities for global text.
topic_local	string		Topic label assigned by BERTopic for the local text.
topic_global	string		Topic label assigned by BERTopic for the global caption.
topic_score_local	float[0..1]	✓	Topic confidence score for the local context.
topic_score_global	float[0..1]	✓	Topic confidence score for the global context.
usage_labels	string[]	✓	Zero-shot usage tags (e.g., *Punchline*, *Joke*).

Table 3: Schema (derived annotations) for MEMEMATCH v1.0.

Data Quality

OCR QA (Local Context) We verified template-string removal via manual audits and spot checks on 50 high-frequency templates. The filter removed recurring artifacts (e.g., watermarks, URLs); residual errors stem from stylized fonts and OCR tokenization splits.

Encoding All textual fields (titles, OCR outputs, captions) are normalized to UTF-8, with standardized whitespace and emojis preserved as Unicode.

Missing Values Policy Fields may be empty if unavailable or extraction failed: `url` (removed posts), `subreddit/title` (API gaps), `topic_*` (`-1` outlier). We do not impute; downstream code should check presence explicitly.

Integrity Constraints For every row, `id` is unique. If `local_text` exists then `sentiment_local[14]` and `topic_local_id` exist; if `global_caption` exists then `sentiment_global[14]` and `topic_global_id` exist.

Types, Encodings, and Controlled Vocabularies

- All text is UTF-8.
- `created_utc` is ISO-8601 (UTC).
- Numeric scores are floats in $[0, 1]$ (up to 16 decimals).
- Integer-like fields (e.g., `score`, `text_length`) may appear as `float64` in CSVs but are integer-valued.
- Topics (`topic_local`, `topic_global`) are string labels; `-1_*` denotes no topic assignment (e.g. `-1_cat_cartoon_comic_dog`).
- Affective columns `sentiment_*[14]` follow a fixed 14-d vocabulary (11 emotions + 3 polarities).
- Array-like fields `usage_labels` stores multiple tags separated by `|` (e.g. `Parody or Spoof|Absurd or Random Humor|Joy or Excitement`).

Implementation note. When loading the dataset, numeric columns should be parsed as floats, with integer-valued fields (e.g., `text_length`, `score`) cast to `int` as needed. Timestamp strings can be converted to Unix epoch seconds if required for temporal analysis.

Exploratory Data Analysis

In this section, we summarize key patterns in MemeMatch, including temporal upload trends, local vs. global differences in sentiment and topic signals, and the overall distribution of predicted meme usage categories.

Temporal Distribution of Meme Uploads

Figure 2 shows a dominant surge in Reddit meme activity: uploads rise sharply in early 2020 and remain elevated through 2021 to early 2022. This period aligns with global COVID-19 lockdowns, when online time, social media engagement, and humor-as-coping increased (Dynel 2021). However, we do not claim that this pattern represents meme culture at large; rather, it reflects patterns within our collected datasets. After

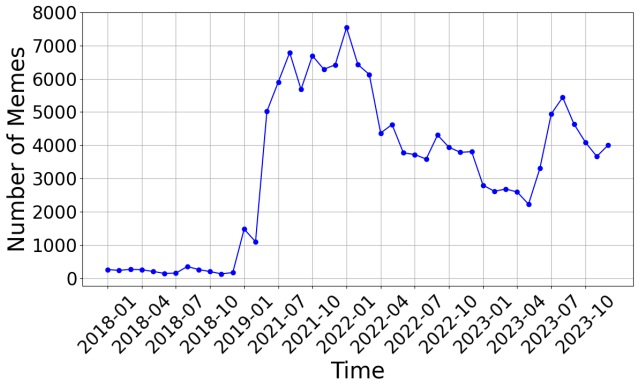


Figure 2: Monthly Reddit meme uploads, with a sharp spike during COVID-19 (2020–2021).

the pandemic peak, uploads decline but remain above prepandemic levels, suggesting that part of the lockdown-driven shift to meme production was retained.

Sentiment Analysis

To compare affective signals from user-added text versus the visual/template description, we analyzed 14-dimensional sentiment–emotion vectors (11 emotions and 3 polarity scores) for both local (OCR + title) and global (BLIP caption) contexts. Figure 3 shows the distributions for all dimensions. A few patterns are evident:

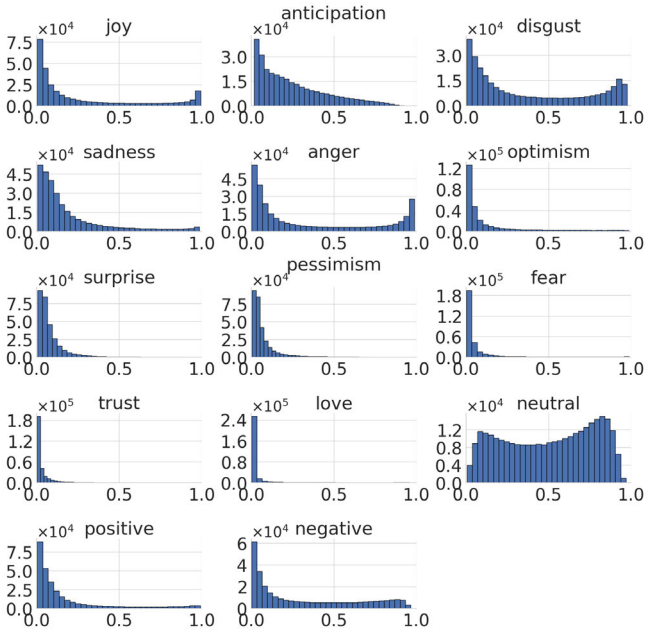
- **Global context is overwhelmingly neutral.** Scores for *neutral_global* cluster near 1.0, while nearly all other global emotions remain close to 0. This reflects that captions of meme templates are descriptive ("a man looking at...") rather than expressive.
- **Local context shows strong emotional variation.** The local distributions are more dispersed, with U-shaped or bimodal patterns for emotions such as *joy*, *anger*, and *disgust*, consistent with prior findings that embedded/overlaid text can strongly shape a meme's expressed emotion, including cases where the same image conveys contrasting emotions depending on the text (Sharma et al. 2024).

These findings reinforce the complementary nature of the dual-context representation: the global layer captures a neutral description of the visual elements, while the local layer carries the expressive and affective information central to meme communication.

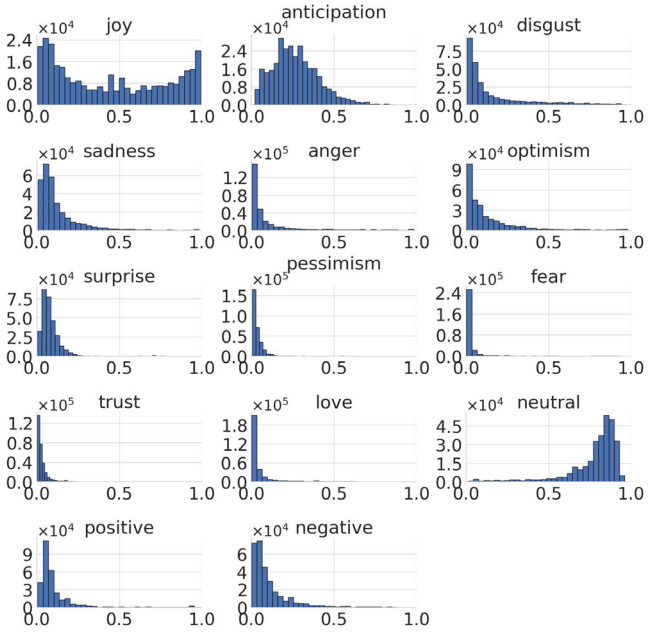
Topic Distribution across Local and Global Contexts

Figure 4 shows the 15 most frequent high-confidence topics (`topic_score_*` > 0.7) from `local_context` and `global_context`. BERTopic labels are keyword summaries; for example, a global-context label such as `2_suit_suits_ties_tie` denotes a broader formal-attire visual theme, not a single exact meme type.

On the **local** side (Figure 4a), topics are more linguistic and reflect meme text, spanning daily situations, online markers, pop-culture references, and event-driven themes. This suggests that the user-edited layer most clearly expresses



(a) Histogram of Sentiment Scores (Local Context).



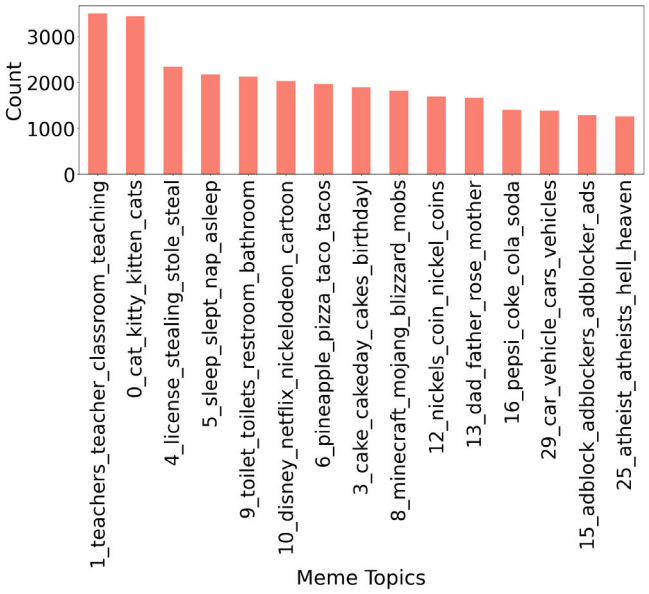
(b) Histogram of Sentiment Scores (Global Context).

Figure 3: Distributions of 14 sentiment and emotion scores for (a) local and (b) global contexts. The x-axis shows sentiment scores, and the y-axis shows meme counts.

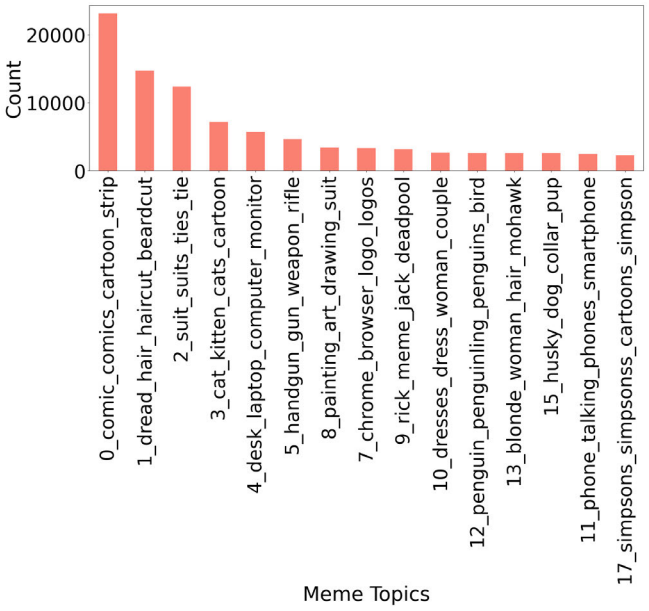
culturally specific and time-sensitive content, consistent with (Barnes et al. 2024).

On the **global** side (Figure 4b), topics look more like visual setups than messages, capturing reusable motifs, appearance cues, common objects/scenes, and animals. In short, the global context captures *visual cues* that make an image recognizable and reusable, while the local context captures

the *message* that changes post to post.



(a) Top 15 High-Confidence Meme Topics in Local Context.



(b) Top 15 High-Confidence Meme Topics in Global Context.

Figure 4: Top 15 most frequent high-confidence meme topics for (a) local context and (b) global context.

Distribution of Meme Usages

Figure 5 summarizes the 15 most common meme usage categories predicted by the zero-shot classifier.

The most frequent categories, *Parody or Spoof*, *Sarcasm or Irony*, and *Absurd or Random Humor*, underscore the central role of humor, exaggeration, and cultural playfulness in meme communication. Mid-frequency categories such as *Reaction or Reply Meme*, *Comparison or Contrast*, and *Wordplay or Pun* suggest that many memes also serve as short-form social responses. Lower-frequency categories, including *Media or Brand Critique*, *Celebrity Mockery*, and *Self-Deprecation*, appear more situational and context-specific.

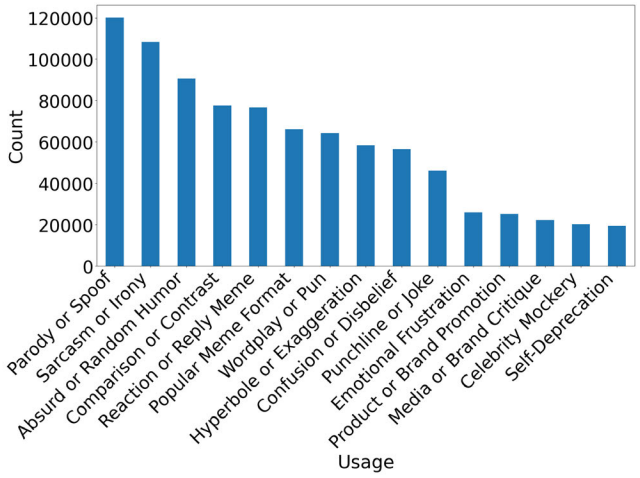


Figure 5: Top 15 most common meme usage categories.

Overall, the usage distribution suggests that memes function primarily as flexible humor mechanisms for self-expression and social interaction, with a smaller fraction serving more specific roles such as critique, mockery, or self-directed humor, in accordance with prior work emphasizing entertainment and self-expression as central motives for meme use (Leiser 2022).

MemeMatch: Context-Aware Multimodal Meme Retrieval

MemeMatch is a context- and intent-aware retrieval system that returns relevant memes or templates from natural-language or image queries [1].

Design and Problem Setup

Goal. Given a natural–language query q or meme image I, return a ranked list of memes or templates that are semantically and usage-relevant.
Representation. Each meme has two complementary textual views: *Local* (user-added text via OCR + title) and *Global* (template caption after masking text).
Principles. (P1) Context-aware (local + global); (P2) Intent-aware (usage labels); (P3) Efficient (precomputed embeddings + cosine search); (P4) Responsible (basic filtering, duplicate suppression).

Case-Based Embeddings

To support diverse query types, MemeMatch precomputes *case-based embeddings*, organized into separate collections based on the information available at query time (Figure 6). Short textual views (local, global, usage) are encoded using `SentenceTransformers("all-mpnet-base-v2")`; relevant views are concatenated in a fixed order (e.g.,

local→usage). Retrieval is cosine similarity between the query vector and the selected collection.

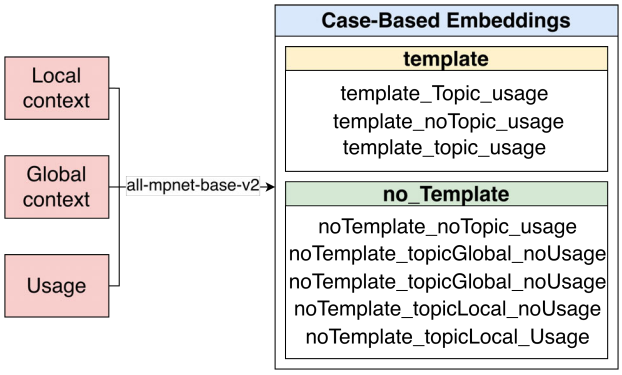


Figure 6: Case-based embedding generation using local, global, and usage text.

Embedding cases. We maintain distinct cases based on (i) search scope (the *template subset* which only contains the meme templates vs. the full meme set) and (ii) which views are present in the query:

- `noTemplate_noTopic_usage`: usage-only intent (e.g., 'a meme for *humor*").
- `noTemplate_topicLocal_noUsage`: topic inferred from local text only (e.g., 'memes about *exams*").
- `noTemplate_topicLocal_usage`: local text + usage intent (e.g., 'memes to *complain* about *dating apps*").
- `noTemplate_topicGlobal_noUsage`: topic from the global caption (e.g., '*SpongeBob* memes").
- `noTemplate_topicGlobal_usage`: global caption + usage intent (e.g., '*SpongeBob* memes to *tease*").
- `template_topic_noUsage`: template subset + topic (e.g., '*Minion* templates").
- `template_topic_usage`: template subset + topic + usage (e.g., '*Minion* templates to *mock* exams").

Queries are then routed to the appropriate collection.

Meme Retrieval using Natural-Language Queries

MemeMatch supports free-form text queries by converting natural language into structured retrieval instructions, as shown in Figure 7.

A lightweight parser built with `GeminiGenerativeAPI` (`gemini-2.5-flash`) extracts three query attributes:

- **Scope** — *templates* vs. full *memes*.
- **Topics** — named entities or concepts (e.g., *Minions*).
- **Usage/Intent** — communicative purpose (e.g., *humor*, *motivation*, *complaint*).

The parser serializes outputs into a compact JSON schema with flags (`need_template`, `has_topics`, `has_usages`) and extracted text fields. The system then:

1. Routes to the appropriate embedding set;
2. Concatenates available views (`topics`, `usages`);
3. Encodes with `SentenceTransformers("all-mpnet-base-v2")`;
4. Retrieves top-N items by cosine similarity.

If neither topics nor intents are detected, MemeMatch falls back to **sentiment-based ranking**: it infers tone (e.g., "funny" → `joy`) and ranks candidates by sentiment scores, defaulting to high-`neutral` when unspecified.

This design lets MemeMatch interpret flexible queries and return memes aligned with semantic content and emotional intent.

Image-Query Meme Similarity Search

Users can upload a meme and retrieve similar instances through two complementary modes (Figure 8):

- **Global context (image-based):** text regions are masked via PaddleOCR; BLIP generates a caption; we perform cosine similarity search in the `noTemplate_topicGlobal_noUsage` collection.
- **Local context (text-based):** overlay text is extracted with EasyOCR and encoded; we perform cosine similarity search in the `noTemplate_topicLocal_noUsage` collection.

Web and Mobile Interfaces

We built two user interfaces on the same backend API:

- **Web app:** text and image retrieval via REST API.
- **Android app:** Kotlin native client with camera uploads, real-time query parsing, Firestore backend, search history, favorites, and social sharing.

Both connect to a FastAPI service managing embedding lookup and fallback ranking.

Applications and Usage

MemeMatch's rich annotations and dual-context embeddings support scholarly and applied work. In the humanities and social sciences, it enables analyses of memes as digital folklore and collective identity, including event-driven commentary. Its sentiment and emotion labels support affective computing and computational humor, consistent with benchmarks such as Memotion (Mishra et al. 2022). At scale (∼301K items), MemeMatch also supports political and social media research by tracing how events (e.g., COVID-19) are reinterpreted via memes (Ortiz et al. 2021) and quantifying trends through retrieval-based aggregation (Joshi, Ilievski, and Luceri 2023).

Practically, MemeMatch can strengthen content moderation by providing training data for multimodal detectors of hate speech, harassment, or misinformation; prior work shows harmful intent often emerges from image–text interactions (Suryawanshi et al. 2020; Kiela et al. 2020). It can also support media literacy education via curated examples and metadata for rhetorical and critical analysis (Silva 2019). The paired retrieval system further supports trend monitoring and cross-platform similarity search using vision–language embeddings (Joshi, Ilievski, and Luceri 2023; Huertas-Tato

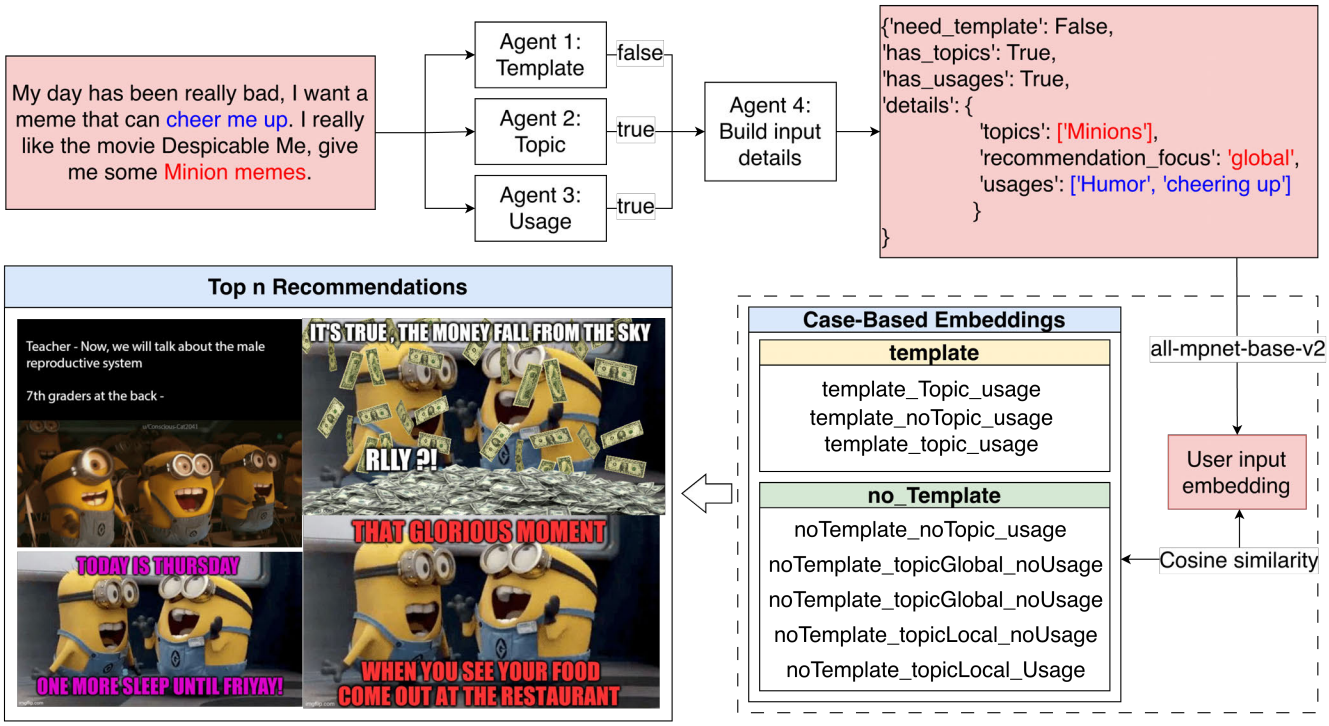


Figure 7: MemeMatch natural–language retrieval pipeline: structured query parsing and embedding routing.

et al. 2024); related efforts such as MemeMind target harmful meme detection at scale (Gu et al. 2025).

MemeMatch is extensible: the pipeline supports swapping in stronger encoders (e.g., fine-tuned CLIP/BLIP-style models), expanding as new templates emerge, linking to external resources such as IMKG (Tommasini, Ilievski, and Wijesiriwardene 2023), and adding new annotation layers (e.g., sentiment taxonomies, humor categories, stance, multilingual coverage). This modular design keeps MemeMatch useful for retrieval and downstream analysis as the field evolves.

FAIR Principles

MemeMatch follows FAIR: **Findable, Accessible, Interoperable,** and **Reusable**.

Findable. Unique IDs link metadata and annotations; schemas (Tables 2–3) document fields and relationships. Hosted on Kaggle with a DOI (https://doi.org/10.34740/kaggle/dsv/14510064) and searchable metadata.

Accessible. Public GitHub repo provides code and artifacts [1]; URLs and template IDs enable lawful access/regeneration from Reddit and ImgFlip. Documentation covers access, structure, and formats.

Interoperable. UTF-8 text, ISO-8601 UTC timestamps, numeric floats/ints; consistent emotion/topic/usage vocabularies. CSV/JSON exports support common ML toolchains (e.g., Pandas, PyTorch, HuggingFace) and multimodal pipelines.

Reusable. GitHub includes pipeline code (OCR, captioning, embeddings, retrieval), `requirements.txt`, setup, and EDA/evaluation notebooks [1]. Reproducing end-to-end results requires underlying dataset access, but the code and documentation remain reusable for other meme collections.

Discussion and Limitations

While MemeMatch advances large-scale meme understanding and retrieval, several limitations remain.

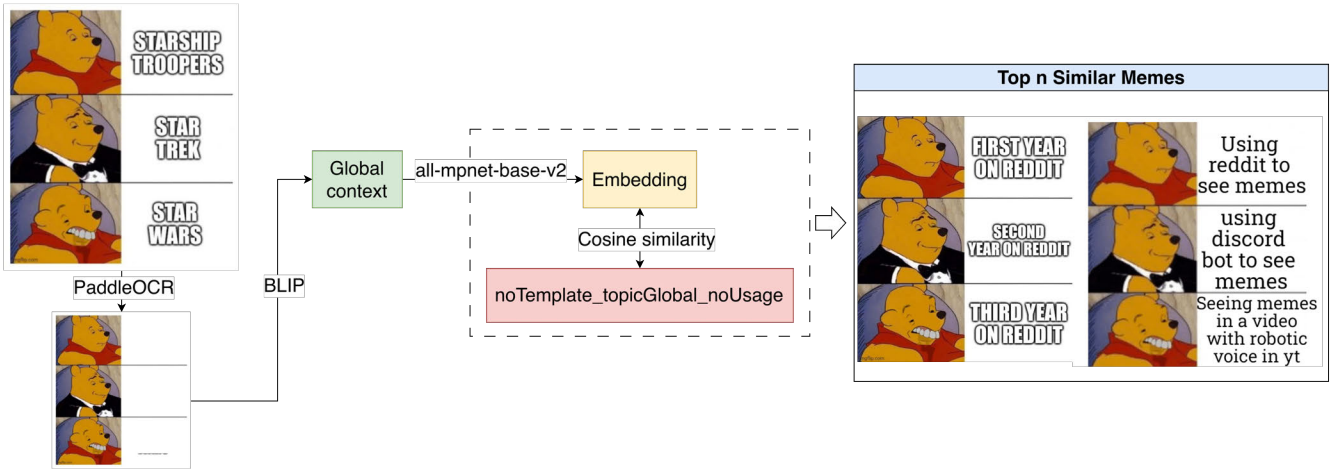
Scope and generalization. MemeMatch covers English memes from Reddit's `r/Memes` and ImgFlip. Retrieval may degrade on unseen styles, templates, or non-English memes without re-embedding or re-indexing.

Temporal coverage. The dataset spans 2018–2023, capturing periods such as COVID-19 but missing earlier and newer meme trends.

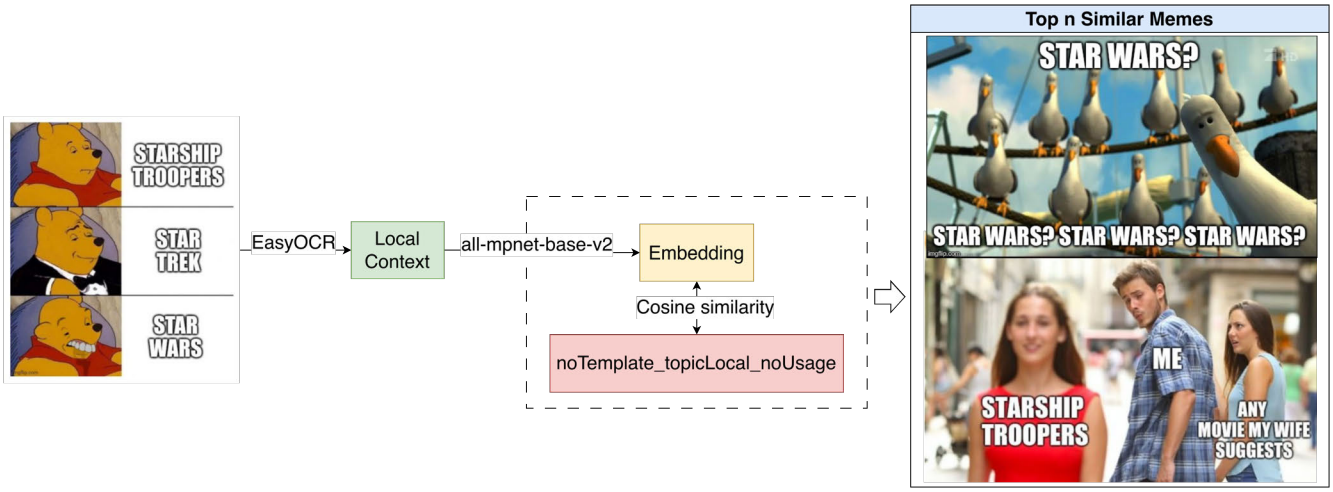
Annotation noise. Automated labels are imperfect. In manual spot checks of 50 random meme samples, OCR and global captions were generally reliable on standard layouts, but failures arose with stylized text, dense overlays, irony, and community-specific meaning. Topics may overlap and zero-shot usage labels are probabilistic, so annotations should be treated as semantic *signals* rather than ground truth.

Biases and implicit meaning. Reddit data and pretrained models may encode societal and linguistic biases, while memes often rely on slang, irony, and subtle image–text cues that generic models may miss.

Sensitive content and misuse. Because memes may contain sensitive content, the dataset should be used with caution. Automated labels are not moderation decisions, and the resource is intended for research and assistive retrieval, not fully automated enforcement.



(a) Global (image) context retrieval.



(b) Local (text) context retrieval.

Figure 8: Dual-mode image-query meme retrieval: (a) global template features, (b) local OCR text.

Despite these constraints, MemeMatch provides a structured multimodal foundation for research on humor, sentiment, and digital communication.

Acknowledgments

DÁK was supported by the Doctoral Students' Excellence Grant Program (DKÖP-25-1-BME-75), while RM was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

Barnes, K.; Juhász, P.; Nagy, M.; and Molontay, R. 2024. Topicality boosts popularity: a comparative analysis of NYT articles and Reddit memes. *Social Network Analysis and Mining*, 14(1): 119.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 830–839.

Boe, B. 2024. PRAW: Python Reddit API Wrapper. https://praw.readthedocs.io/. Accessed: 2025-05-05.

Camacho-collados, J.; Rezaee, K.; Riahi, T.; Ushio, A.; Loureiro, D.; Antypas, D.; Boisson, J.; Espinosa Anke, L.; Liu, F.; and Martínez Cámara, E. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In Che, W.; and Shutova, E., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–49. Abu Dhabi, UAE: Association for Computational Linguistics.

Du, Y.; Li, J.; Yan, R.; et al. 2020. PaddleOCR: An Ultra Light-weight OCR System. https://github.com/PaddlePaddle/PaddleOCR. Accessed: 2025-05-05.

Dubey, A.; Moro, E.; Cebrian, M.; and Rahwan, I. 2018. MemeSequencer: Sparse Matching for Embedding Image

Macros. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, 1225–1235. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450356398.

Dynel, M. 2021. COVID-19 memes going viral: On the multiple multimodal voices behind face masks. *Discourse & Society*, 32(2): 175–195.

Grootendorst, M. 2020. KeyBERT: Minimal keyword extraction with BERT.

Grootendorst, M. 2022. BERTopic: Neural Topic Modeling with Class-based TF-IDF. https://maartengr.github.io/BERTopic/. Accessed: 2025-05-05.

Gu, H.; Yu, Q.; Hou, S.; Fang, Z.; Wu, H.; and He, Z. 2025. MemeMind: A Large-Scale Multimodal Dataset with Chain-of-Thought Reasoning for Harmful Meme Detection. In *Proc. 33rd ACM Int. Conf. on Multimedia (MM)*.

Huertas-Tato, J.; Koutlis, C.; Papadopoulos, S.; Camacho, D.; and Kompatsiaris, I. 2024. A CLIP-based Siamese Approach for Meme Classification. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Jaided AI. 2020. EasyOCR: Ready-to-use OCR with 80+ supported languages and all popular writing scripts. https://github.com/JaidedAI/EasyOCR. Accessed: 2025-05-05.

Joshi, S.; Ilievski, F.; and Luceri, L. 2023. Contextualizing Internet Memes Across Social Media Platforms. *arXiv e-prints*, arXiv:2311.11157.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Klinger, E.; and Starkweather, D. 2008. pHash: The Open Source Perceptual Hash Library. http://www.phash.org. Accessed: 2024-04-25.

Leiser, A. 2022. Psychological Perspectives on Participatory Culture: Core Motives for the Use of Political Internet Memes. *Journal of Social and Political Psychology*, 10(1): Article e6377.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Loureiro, D.; Barbieri, F.; Neves, L.; Espinosa Anke, L.; and Camacho-collados, J. 2022. TimeLMs: Diachronic Language Models from Twitter. In Basile, V.; Kozareva, Z.; and Stajner, S., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 251–260. Dublin, Ireland: Association for Computational Linguistics.

Ludwig, M. 2024. PMAW: Pushshift Multithreaded API Wrapper. https://github.com/mattpodolak/pmaw. Accessed: 2025-05-05.

McInnes, L.; Healy, J.; and Astels, S. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11): 205.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Mishra, S.; Suryavardan, S.; Patwa, P.; Chakraborty, M.; Rani, A.; Reganti, A.; Chadha, A.; Das, A.; and Sheth, A. 2022. Memotion 3: Dataset on Sentiment and Emotion Analysis of Code-mixed Hindi-English Memes. In *Proceedings of the First Workshop on Sense and Sentiment in Memes (SenseMeme)*, volume 3555 of *CEUR Workshop Proceedings*.

Ortiz, J. A. F.; Corrada, M. A. S.; Lopez, E.; and Dones, V. 2021. Analysis of the use of memes as an exponent of collective coping during COVID-19 in Puerto Rico. *Media International Australia*, 178(1): 168–181.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, 8748–8763. PMLR.

Richardson, L. 2013. Beautiful Soup Documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Accessed: 2025-05-05.

Sanderson, B.; and Rigby, M. 2013. We've Reddit, have you?: What librarians can learn from a site full of memes. *College & Research Libraries News*, 74(10): 518–521.

Sharma, D.; Bhageria, D.; and Das, A. 2020. SemEval-2020 Task 8: Memotion Analysis—The Visuo-Lingual Metaphor! In *Proceedings of the 14th Workshop on Semantic Evaluation (SemEval 2020)*, 759–773. Barcelona, Spain (Online): Association for Computational Linguistics.

Sharma, S.; S, R.; Akhtar, M. S.; and Chakraborty, T. 2024. Emotion-Aware Multimodal Fusion for Meme Emotion Detection. *IEEE Transactions on Affective Computing*, 15(3): 1800–1811.

Silva, L. 2019. To Meme or Not to Meme? Using Memes to Teach Media Literacy Skills. KQED Education.

Suryawanshi, S.; Chakravarthi, B. R.; Arcan, M.; and Buitelaar, P. 2020. Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. In Kumar, R.; Ojha, A. K.; Lahiri, B.; Zampieri, M.; Malmasi, S.; Murdock, V.; and Kadar, D., eds., *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 32–41. Marseille, France: European Language Resources Association (ELRA). ISBN 979-10-95546-56-6.

Tommasini, R.; Ilievski, F.; and Wijesiriwardene, T. 2023. IMKG: The Internet Meme Knowledge Graph. In Pesquita, C.; Jimenez-Ruiz, E.; McCusker, J.; Faria, D.; Dragoni, M.; Dimou, A.; Troncy, R.; and Hertling, S., eds., *The Semantic Web*, 354–371. Cham: Springer Nature Switzerland. ISBN 978-3-031-33455-9.

Wang, B.; Lin, J.; Bai, Z.; Wang, Z.; Sun, S.; Jin, Z.; Wen, Z.; Tu, G.; Li, J.; Cambria, E.; and Xu, R. 2025. Internet Meme on Social Media: A Comprehensive Review and New Perspectives. *Information Fusion*, 104102.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts? **Yes**
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes**
   (e) Did you describe the limitations of your work? **Yes**
   (f) Did you discuss any potential negative societal impacts of your work? **Yes**
   (g) Did you discuss any potential misuse of your work? **Yes**
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research? **Yes**
   (i) Have you read the ethics review guidelines and ensured 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? **NA**
   (b) Have you provided justifications for all theoretical results? **NA**
   (c) Did you discuss competing hypotheses or theories? **NA**
   (d) Have you considered alternative mechanisms or explanations? **NA**
   (e) Did you address potential biases or limitations in your theoretical framework? **NA**
   (f) Have you related your theoretical results to the existing literature? **NA**
   (g) Did you discuss the implications of your theoretical results? **NA**
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? **NA**
   (b) Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results? **Yes**
   (b) Did you specify all the training details? **Yes**
   (c) Did you report error bars? **NA**
   (d) Did you include the total compute and resource types used? **Yes**
   (e) Do you justify how the proposed evaluation is sufficient and appropriate? **Yes**
   (f) Do you discuss the cost of misclassification and fault tolerance? **Yes**
5. Additionally, if you are using existing assets or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **Yes**
   (b) Did you mention the license of the assets? **Yes**
   (c) Did you include any new assets in the supplemental material or as a URL? **Yes**
   (d) Did you discuss whether and how consent was obtained? **Yes**
   (e) Did you discuss whether the data contains personally identifiable information or offensive content? **Yes**
   (f) If curating new datasets, did you discuss how they are FAIR? **Yes**
   (g) If curating new datasets, did you create a Datasheet for the Dataset? **Yes**
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions and screenshots? **NA**
   (b) Did you describe participant risks or IRB approvals? **NA**
   (c) Did you include estimated hourly wages and compensation? **NA**
   (d) Did you discuss how data is stored, shared, and deidentified? **NA**