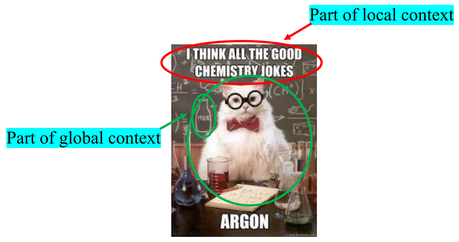
MemeMatch: A Large-Scale Dual-Context Multimodal Dataset and Retrieval System for Internet Memes

Tri An Le and Dr. Qixin Deng

Department of Mathematics and Computer Science, Wabash College

Introduction

MemeMatch is a **large-scale dual-context meme dataset + retrieval system** that separates **local user text** from **global template semantics**, adds rich annotations (emotion, topic, usage intent), and enables **context-aware search** from text or images.



1. **Local context = what the meme says** (captures the **user-added overlay text** and the **situational message)**
2. **Global context = what the meme shows** (captures the **underlying image/template semantics** without the added text)

Problem

Meme meaning comes from **both** the template image and the **user-added text**, so the same template can convey different messages, making search and analysis unreliable with a single representation.

Contributions

1. **Large-scale dataset:** ~301K memes and 2,083 templates collected from **Reddit** (organic social memes) and **ImgFlip** (template-based memes).
2. **Rich annotations: 14D sentiment/emotion** (11 emotions + 3 polarities), **BERTopic** (300 local topics, 200 global topics), and **28 usage-intent labels**.
3. **Dual-context design:** separates **Local** (overlay text + title) vs **Global** (template caption) to preserve both **message** and **template meaning.**
4. **Context-aware retrieval system:** supports **text + image queries** using **case-based embeddings** and an **LLM-based query parser.**
5. **Large-scale EDA findings:** reveal trends in **emotion, topics,** and **usage** over time, and their association with **engagement/virality patterns.**

Dataset Structure

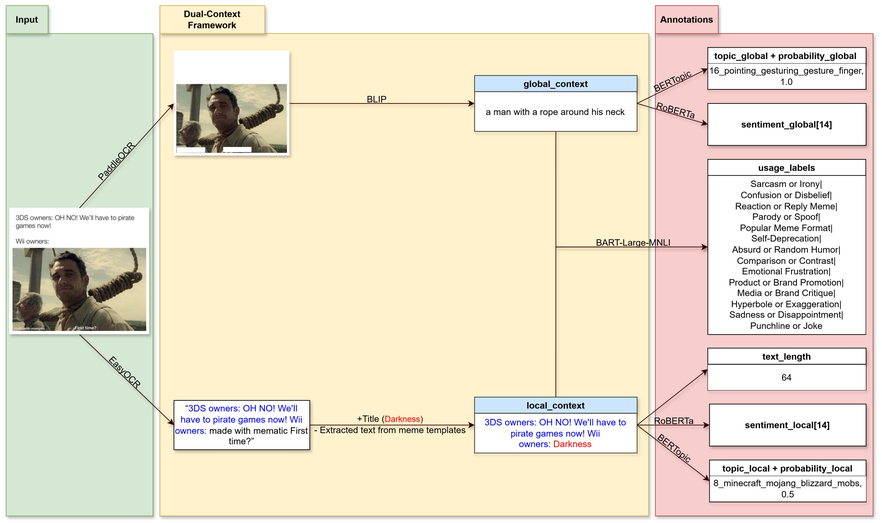
Field	Type	Req.	Description
filename	string	✅	Image filename, used as the unique identifier across all tables. For ImgFlip memes, the filename also indicates the template name.
created_utc	string		Reddit meme upload timestamp (ISO-8601 format).
score	int		Reddit upvote count at crawl time.

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.

Schema (derived annotations) for MemeMatch v1.0.

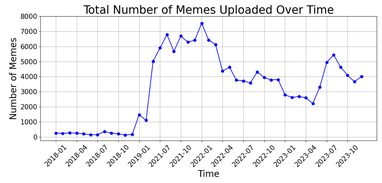
Methodology



1. Data Collection and Preprocessing (Input Stage)

1.1. ImgFlip: 3,127 templates (Apr 2025) → **2,083 unique** after dedup; **153,792 template-labeled memes** (≈ **74 per template**).
1.2. Reddit: Jan 2018–Dec 2023, up to **1,000 posts/day → 899,522 images** with metadata; upvotes reflect the **Jan 2024 crawl window** (not lifetime totals).
1.3. Deduplication: Removed duplicates using **perceptual hashing (pHash).**
1.4. Final combined corpus

- **146,991 Reddit memes** (no template labels).
- **153,792 ImgFlip memes** across **2,083 templates.**
- Combined dataset is the input to the **Dual-Context Framework.**



2. Dual-Context Framework Stage

2.1. Local Context Extraction (Message-Level)

- **EasyOCR** extracts overlay text; remove common template artifacts by filtering strings found in **2,083 templates** (e.g., watermarks/URLs), then append **Reddit title → local_context.**

2.2. Global Context Extraction (Template-Level)

- **PaddleOCR** locates text to mask; **BLIP** captions the masked base image → **global_context** *(for Reddit memes, "template" = base image after masking).*

2.3. Why Hybrid OCR?

- **EasyOCR** is strong for fast text extraction at scale, while **PaddleOCR** provides tighter boxes for masking before captioning.

2.4. Output: Each meme produces a paired representation: **local_context + global_context,** which are used in downstream **annotation and retrieval (Annotations Stage).**

3. Annotations Stage

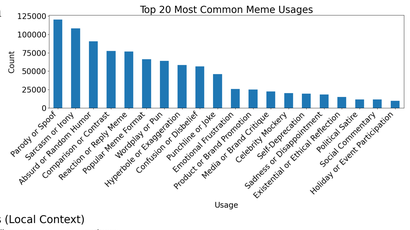
3.1. Sentiment & Emotion (How the meme feels)
We apply two **RoBERTa-based** models (CardiffNLP / TweetNLP) to **both local and global contexts**:

- **Emotion model:** predicts **11 emotions** (*Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, Sadness, Surprise, Trust).*
- **Sentiment model:** predicts **3 polarities** (*Positive, Neutral, Negative).*

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

3.2. Usage Intent Labels (How the meme is used)

- We infer communicative intent using **zero-shot classification** (BART-MNLI) over **28 predefined usage labels** on concatenated **local + global** text; keep labels with **confidence ≥ 0.70**.

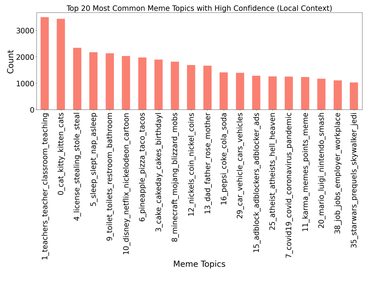


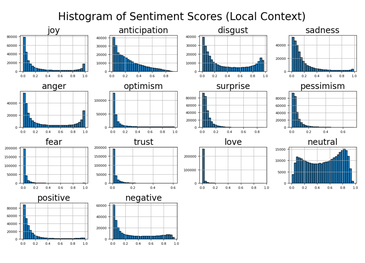
3.3. Topic Modeling (What the meme is about)

- Cluster themes in both contexts: **300 local topics** and **200 global topics**; each meme gets a **topic ID + probability**, with low-confidence items marked as **outliers (-1)**.

3.4. Auxiliary Feature

- **Text length:** character count of the **local context** (text_length), which complements higher-level annotations and supports downstream analysis.





MemeMatch: Context-Aware Multimodal Meme Retrieval

Goal: given a **text query *q*** or **meme image *I***, return a ranked list that matches:

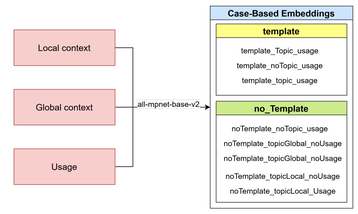
- **what the meme is about** (topic/meaning), and
- **how the user wants to use it** (intent, e.g., joke, complain, motivate).

1. Design Principles

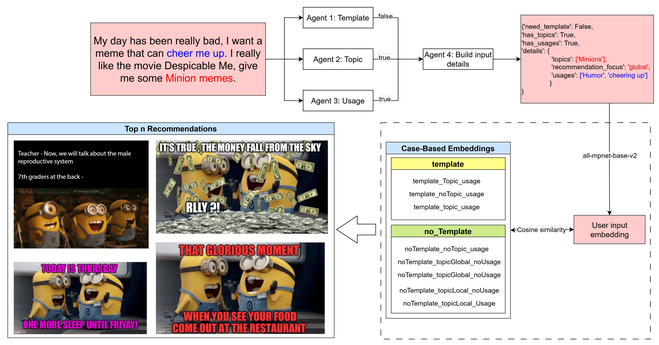
- **Context-aware:** uses both **Local** (overlay text + title) and **Global** (template caption) views.
- **Intent-aware:** incorporates **usage labels** (e.g., sarcasm, parody).
- **Efficient:** fast retrieval using **precomputed embeddings + cosine similarity.**
- **Responsible:** basic filtering and duplicate suppression.

2. Case-Based Embeddings (Fast Retrieval Backbone)

- Different queries contain different information (topic vs intent vs template) so MemeMatch precomputes multiple **embedding indexes ("cases")** so it can retrieve efficiently for each query type.
- **Encoder:** SentenceTransformers all-mpnet-base-v2.
- Retrieval uses **cosine similarity** between the query embedding and the selected index.



3. Natural-Language Retrieval (Text Queries)



Users enter free-form text queries (e.g., "a funny meme about finals"); an LLM-based query parser extracts:

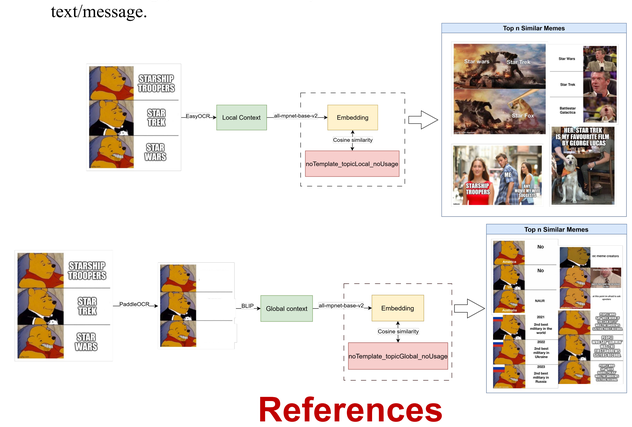
- **scope** (templates vs memes),
- **topic** (e.g., Minions/exams),
- **intent** (e.g., humor/motivation/complaint),

then routes to the right embedding index and retrieves top-k results by similarity.
Fallback: If topic/intent is unclear, MemeMatch falls back to **tone matching** (e.g., "funny" → joyful memes) using sentiment scores.

3. Image-Query Retrieval (Upload a Meme)

Users can upload an image and retrieve similar memes in **two complementary modes**:

- **Global (template) mode:** mask text → caption underlying image → retrieve memes with similar template meaning.
- **Local (message) mode:** read overlay text → retrieve memes with similar text/message.



References

1) **Le, D.T.A., Koller, D. A., Deng, Q., and Molontay, R. "MemeMatch: A Large-Scale Dual-Context Multimodal Dataset and Retrieval System for Internet Memes." Under review at ICWSM 2026.**