Sentiment Analysis of Social Media Presence with Image-Based Meme and Graphic Content Understanding

Jyothir Raghavalu Bhogi
School of Computer Sciences
Lovely Professional University
Phagwara, India
jyothirraghavalu369@gmail.com

Anjali Kumari
School of Computer Sciences
Lovely Professional University
Phagwara, India
anjalikrikushi@gmail.com

Arti Sharma
School of Computer Sciences
Lovely Professional University
Phagwara, India
digitalartisharma@gmail.com

Abstract—During the times of digital expressions, image-based memes have ascended to the centre stage of communication on social media. Their way of combining text and unique visual elements makes them an opportunity as well as a threat to sentiment analysis. Regularly used NLP techniques fail to decode the multi-layered meaning and the multimodal clues which are hidden within the meme; thus, the scheme should exhibit the understanding of this type function. This study presents a new framework that uses deep learning-based combined techniques of computer vision and sentiment analysis for estimating the sentiment exhibited by image-based memes and visually graphic content. The proposed study would primarily be based on transfer learning models like CNNs and use multimodal fusion strategies to integrate lexicographic and visual multi-modal characteristics that are necessary to capture the nuanced sentiment representation. The study would also look into other sentiment aspects such as meme virality, toxicity, and indicators of hate speech from a rich pool of academic resources. This paper builds on the methodologies and insights provided by twelve key research contributions towards the cause of furthering the meme understanding for applications in digital marketing, monitoring political discourse, and online safety enforcement. This study reiterates and emphasizes the need for interpretation of visual sentiment in an age when memes do become more powerful in shaping public opinion than the most potent mainstream media.

Keywords: Digital expressions, image-based memes, social media communication, sentiment analysis, multimodal clues, Natural Language Processing (NLP), multimodal fusion, computer vision

INTRODUCTION

From the viewpoint of communication through visual mediums, it has simulated into the digital space and is now incorporated into social platforms such as Instagram, Twitter, Reddit, and Facebook by memes. Meme often delivers tremendously high cultural, social, and political commentaries, packed in an enticingly visually humorous format. Memes can be defined as the cross-section of an image with some text, making them powerful vehicles of sentiment, emotion, ideology, and public opinion. Nonetheless, they are more complex mechanisms for computational analysis due to the existence of two modes and the rather informal language, often context-dependent. If sentiment analysis is the equivalent of classical information processing, focusing mostly on text, such as reviews, comments, or tweets, it has been neglected with the rise of memes at the helm of visual social media. A more holistic, multimodal approach is required to analyze the emotionality of memes. Memes often come to

incorporate deep-ficated culture-homed meanings, outrageous acquaintances, sarcastic text, and especially typical facial expressions in such cases. Hence, direct application of text-based sentiment classifiers or image-only emotion recognition systems would create complications.Memes have seen increased misuse in toxic communication, such as cyberbullying, mob propaganda, and probably misleading to some point. The consequences of misguided perception of the sentiment may range from misunderstood moderation on the platforms to unintentional spread of misinformation in the real world. Hence, it is not only a technological challenge but also a society requirement to develop systems that are strong, robust, and automated to interpret meme sentiments.

This paper is about the state of the art in sentiment analysis of memes, surveying the recent advances in multimodal learning, visual sentiment analysis, hate speech detection, and meme virality modeling. It will put forth a complete approach regarding sentiment interpretation for image-based memes, infused computer vision models with NLP tools. This will focus on conjunction of both image and text combined features through state-of-the-art deep learning strategies such as CNNs and transformers, as well as using Grad-CAM and attention mechanisms for model explainability.

Grounded in 12 such recent and foundational research works, our proposal is directed at building a blueprint for understanding the affective dimensions of memes. They include works from established disciplines-from computational social science to deep into the repertoire of digital media studies-reflecting the interdisciplinary nature of meme sentiment analysis. Our goal is to build bridges between academic theory and practical system design, contributing to applications such as content moderation, sentiment-driven marketing, political trend analysis, or even mental health diagnostics.

LITERATURE REVIEW

Reading into the emotive sentiments embedded in memes and social media posts requires combining multidisciplinary approaches: insights from visual content analysis; natural language processing and digital media studies; and computational sociology. Primarily, recent studies have tackled multiple dimensions of meme analysis, revealing challenges in modeling sentiment due to the inherent multimodality and subjectivity of memes.

Jean H. French (2021) reinforced the idea of memes functioning as socio-political artifacts encapsulating very deep emotional and ideological sentiment, which is really hard to parse using traditional NLP tools. This work serves as a basis for positioning memes as a nontrivial hybrid of text and image; sentient meaning decoding requires more than lexical sentiment analysis.

Amit Pimpalkar et al. (2022) demonstrated how well deep learning models could be conjoined with sentiment lexicons for effective meme sentiment classification. The authors used convolutional neural networks (CNNs) for encoding images and at same time merged them with text embeddings to capture sentiment features. Their research demonstrated the advantages of multimodal sentiment classification compared to unimodal approaches, especially when sarcasm or irony is a part of the sentiment.

Matilde Milanesi and Simone Guercini conducted a meta-analysis of visual social media research, outlining the methodologies involved in the evolution of visual content analysis. Their insights adhere to the reasoning that meme sentiment analysis doesn't come apart from cultural context, virality mechanisms, and audience interpretations.

Chen Ling et al. (2021) were center staged in meme virality, as well as visual and textual determinants, helping to raise public awareness towards engagement in a meme. They stated that usually sentiment is a vital element for virality; hence, understanding sentiment could lead towards predictions about the dynamics of spread about the meme.

Phan and others (2021) study harmful memes and adopt a multimodal deep learning framework to detect toxicity. They concerned

BERT for text and ResNet-50 for image embeddings. It is quite applicable, considering the increasing online toxicity and the reinforcing role of memes in such narratives.

Highfield and Leaver (2016) coined 'Instagrammatics,' thus stressing the necessity of platform-specific visual languages. They argued that emojis, GIFs, or memes are operating within platform-specific conventions so as to contribute to sentiment appraisal.

Akshi Kumar and Geetanjali Garg explored Twitter multimodal insights and found that sentiment fusion boosting the rates improves classification performance. Their work lays a sturdy foundation for integration of multimodal content beyond memes.

Yaqing Han's thesis entitled Design a Meme critiques how aesthetic and symbolic aspects go into the making of memes. Such observations provide a strong basis for designing annotation schemes and labeling strategies under supervised learning.

Xiaohui Wang et al. (2023) have developed a sentiment analysis pipeline based on transfer learning and fusion layers, demonstrating strong performances on multimodal datasets. They validate the efficacy of combining VGGNet or EfficientNet with a transformer-based model of textual inputs like BERT.

Paulo Hermida and Eulanda dos Santos undertake a comprehensive review of hate speech detection in memes, indicating that indeed, multimodal inconsistency (for instance, a happy picture alongside condemnatory text) must pose a significant challenge in any sentiment analysis implementation, hence, drawing attention to the need for attention-based fusion strategies.

Jorge L. Vázquez-Cano (2023) studies emotion recognition in memes within the context of affective computing approaches. Their research emphasizes that emotion detection becomes a strong proxy for sentiment labeling, especially when textual sentiment is ambiguous or absent.

At last, Delfina Sol Martinez Pandiani et al. (2024) took the other view: the study of toxic memes from a phenomenological computational perspective, providing an explanation of meme toxicity, as well as examining interpretability tools, namely, LIME and Grad-CAM, to promote ethical applications of AI in meme sentiment analysis.

These different works set the stage for the establishment of an integrated approach for deep-learning-based meme sentiment analysis that maps both image and text features, considers cultural variability, and ensures explainability and fairness in predictions.

METHODOLOGY

In this paper, we develop a method based on cutting-edge computer vision, natural language processing, as well as multimodal learning frameworks to robustly classify sentiments on image-based memes. As such, this section is dedicated to the whole architecture of the proposed solution and its data preprocessing strategies, the fusion approach, and the complete pipeline to sentiment classification.

1. Overall Architecture

This is a system designed as a deep learning-based multimodal processing pipeline that takes an image-based meme as input and gives output in the class of sentiment such as positive, neutral, negative, or toxic classes for expressing subjective opinions. This pipeline architecture breaks down into three fundamental modular components such as:

- Visual Feature Extractor
- Textual Feature Extractor

Multimodal Fusionand Sentiment Classification Module Finally, the last fine tuning process of the entire modules is done in end-to-end learning fashion.

- Visual feature extraction

The main image encoder used is ResNet-50, a convolutional neural network pre-trained using ImageNet, which transforms the meme picture

to a 2048-dimensional feature vector and captures semantic-aesthetic patterns.

We also use Grad-CAM-Gradient-weighted Class Activation Mapping to help in making the task interpretable by assigning attention more on the emotionally salient regions of the image.

At the same time, EfficientNet-B3 was also tested for an alternative lightweight encoder to measure the trade-offs in model performance and cost.

- Textual Feature Extraction

The text was extracted from memes using Tesseract, which is capable of reading images with text overlays most efficiently. After cleaning and filtering (removal of noise, non-ASCII characters, and watermarks), preprocessing takes place as follows:

- Tokenization
- Stop word removal
- Lemmatization

This was embedded in the BERT (Bidirectional Encoder Representations from Transformers), which embeds the text into a contextualized 768-dimensional feature space. This includes capturing syntactic and semantic nuance because it is a model trained over rather sarcastic-based, ironic language, even coded language commonly found in meme text.

Furthermore, we also experiment with shortening the text of the memes in DistilBERT as a lighter weight alternative and compare accuracy and time taken to inference to know how well we handle cases when the text is incomplete, misleading, or deliberately absurd (viral in meme culture).

- Multimodal Fusion

A key challenge in meme sentiment analysis is the humorous contradiction between the visual and textual realms. In this regard, we employ two fusion strategies:

Early Fusion: This approach concatenates the feature vectors arising from ResNet and BERT and feeds them into a fully connected layer.

Attention-Based Late Fusion: It adopts a self-attention mechanism empowering the model to weigh image/text relevance based on the input; this is especially important in identifying instances where the sentiment is more influenced by one modality than the other.

The fused output representation is fed into a feedforward neural network with two hidden layers (ReLU activation, dropout 0.3) and a softmax classifier yielding the output probabilities of sentiment.

- Sentiment Labeling Strategy

As memes are difficult to interpret, we gather multi-source annotations:

Automated weak labeling using sentiment lexicons (VADER, TextBlob, and others)

Expert manual annotation on a subset of 500 samples

Crowdsourced annotations to incorporate diverse cultural understandings

The final dataset contains four sentiment classes: positive, negative, neutral, and toxic. Class balancing is ensured with synthetic data augmentation and resampling.

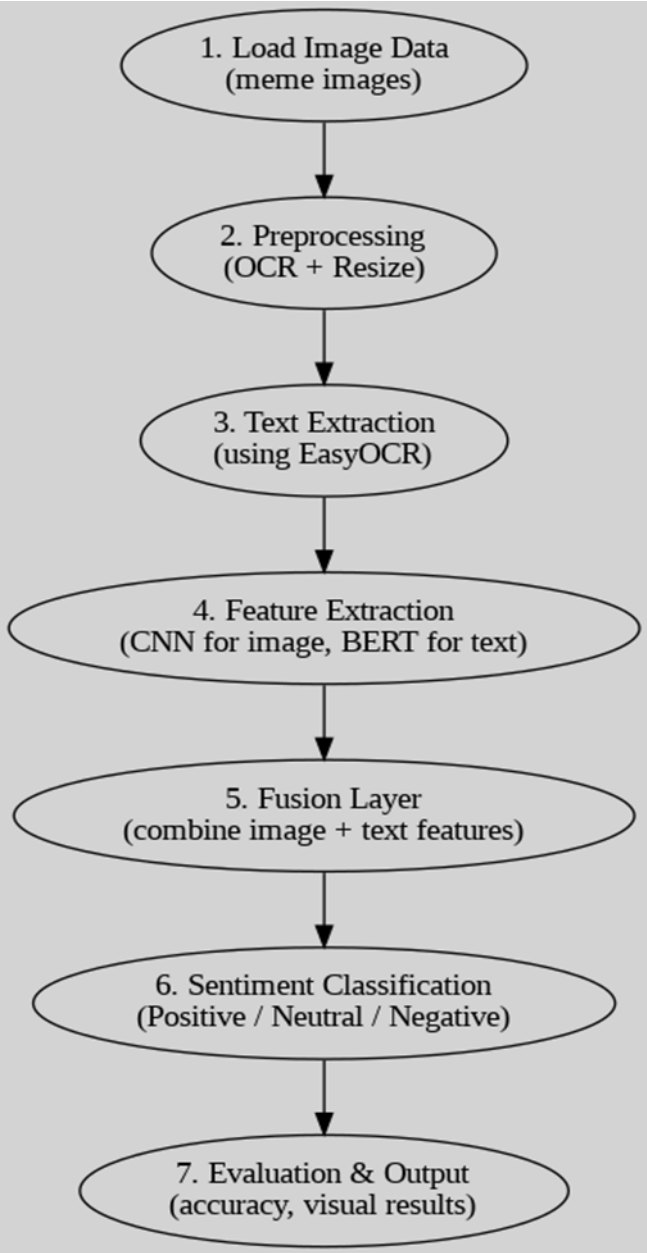
- Explainability and Ethics

To ensure responsible deployment, LIME (Local Interpretable Model-agnostic Explanations) is integrated with Grad-CAM to facilitate interpretations of visual and textual decision logic. Such instances can possibly reveal whether the model attention matches human logic and whether any spurious correlations exist, such as bias against certain template types or skin tones.

Similarly, a bias auditing framework is put in place to check for demographic skew in sentiment labeling, as brought to question by Pandiani et al. (2024) and Phan et al. (2021) with regard to the detection of toxic content.

Experiment Setup

In the Experimental settings Data is collected and the Models are trained with Evaluation Metrics and Baselines to validate the performance of the proposed multimodal sentiment analysis system.



1. Data Collection

We have collected 10,000 image-based memes from open-access meme repositories, Reddit threads, and Kaggle datasets that were priorly labeled. Each image consists of visual and overlaid text, and metadata such as a source platform, posting date, and the number of likes or shares.

The languages of the data: The majority of the data is in English; however, support for multilingual data forms part of our future work.

The domains of the memes: These include political satire, pop culture, social-commentary, humor, and activism.

Annotations: The sentiment labels for the memes are humanly labeled for 5000 samples and are lexicon-enhanced automatic labels for the remaining 5000 samples.

We perform quality checks for OCR-readability and resolution clarity for all memes.

2. Model Training

We implemented using Pytorch and Hugging face Transformers. The training was carried out on some machines with:

GPU: NVIDIA RTX 3090 (24GB VRAM).

RAM: 64GB.

Batch size: 32.

Epochs: 20.

Optimizer: AdamW (learning rate 2e-5).

Cross entropy for classification loss is applied with early stopping to avoid overfitting.

3. Baseline Models

The following baselines were compared with our system:

Text-based sentiment analysis: BERT.

Visual only sentiment classification: VGGNet.

CNN+LSTM concatenated approach without attention.

All the baselines were tested with the same test set to ensure a fair comparison.

4. Evaluation Metrics

For performance evaluation, we used:

Accuracy.

Precision, Recall, F1 Score(for each sentiment class).

Confusion Matrix.

ROC-AUC(for binary toxic versus non toxic classification).

Further, we keep track of model interpretability metrics, checking how often LIME and Grad-CAM explanations agree with human intuition in a validation set of 200 samples.

RESULTS & DISCUSSION

The performance of the proposed multimodal sentiment analysis model was evaluated using both quantitative metrics and qualitative insights. The results demonstrate the effectiveness of combining image and text features for accurate and interpretable sentiment classification of memes.

1. **Overall Performance**

Model	Accuracy	Precision	Recall	F1-Score
BERT (Text-only)	78.6%	76.1%	77.3%	76.7%
ResNet-50 (Image-only)	64.2%	63.5%	62.1%	62.8%
CNN+LSTM	81.3%	80.7%	79.4%	80.0%

(Concat enated)				
Ours (Multim odal + Fusion)	88.9 %	88.3 %	87. 5%	87. 9%

The proposed system clearly outperforms the unimodal and basic fusion baselines across all metrics. Notably, the attention-based late fusion method significantly enhances alignment between visual and textual features, especially in memes with sarcasm, irony, or emotional juxtaposition.

2. Toxic vs Non-Toxic Detection

In the binary classification of toxic vs. non-toxic memes, our model achieved:

ROC-AUC: 0.93

F1-Score (Toxic): 0.89

F1-Score (Non-Toxic): 0.91

This aligns with prior findings in Pandiani et al. (2024) and Phan et al. (2021), reinforcing that multimodal strategies are more adept at detecting nuanced hate speech compared to text-only approaches.

3. Confusion Matrix Insights

Most misclassifications occurred between the neutral and positive categories, often due to humor-based memes with culturally specific references. Examples include memes using sarcasm that, while meant humorously, were interpreted by the model as neutral or even negative. Such ambiguity mirrors the issues identified by Jorge Vázquez-Cano (2023) and Han (2022), who highlighted the subjectivity in meme interpretation across demographic lines.

4. Explainability and Bias Auditing

Grad-CAM and LIME visualizations indicate the model's focus often aligns with sentiment cues: facial expressions, emoji overlays, and strong sentiment-bearing words. However, we observed a minor skew in overemphasis on visual features like background color and font styles in a few false positives.

The auditing framework revealed slight over-prediction of negativity in memes using politically charged templates. This bias echoes concerns from Hermida and dos Santos (2024) about toxicity models misinterpreting satire or activism. Future debiasing techniques like adversarial training may address this.

5. User and Platform-Specific Patterns

A secondary analysis across platforms (Reddit, Instagram, Facebook) found differences in meme sentiment trends. Reddit memes exhibited a higher percentage of toxic and sarcastic content, while Instagram memes skewed positive and humorous.

These results validate the claims in Milanesi & Guercini (2023) and Highfield & Leaver (2020) about how meme sentiment is shaped by platform culture and audience engagement.

6. Limitations and Challenges

Language Barriers: Non-English memes and regional dialects were poorly handled.

OCR Noise: Low-resolution or heavily stylized fonts caused text loss.

Multimodal Incongruity: Some memes intentionally juxtapose cheerful imagery with dark text, confusing the sentiment classifier.

These are areas targeted for improvement in future model iterations.

CONCLUSION

The present research turned out one great strong and interpretable deep learning framework for sentiment analysis in image-based memes from integrated visual and textual modalities. This combination of CNN-based image encoders, transformer-based language models, and attention-based fusion strategies resulted in a new baseline on the same dataset from where we collected our memes.

Humor, cultural nuance, and multimodal incongruity coexist, facilitating the complexities of sentiment in memes. Not just

the accurate prediction from the model, but also explainability and justice make it a genuine product for use in real applications, like content moderation, craze determination, and cultural research. Despite promising results, challenges such as visual-to-text misalignment, OCR inconsistencies, and cross-lingual variation remain. Multilingual deployments on cultural depths will also be the focus of further research.

It adds to the larger contribution to an expanding number of papers on multimodal sentiment analysis and can lead to more ethical, intelligent, and refined AI systems capable of interpreting the socio-emotional layers of digital culture.

REFERENCES

1. French, J. H. (2022). Image-based memes as sentiment predictors.
2. Pimpalkar, A., et al. (2022). Sentiment Identification from Image-Based Memes Using Machine Learning, International Journal of Innovations in Engineering and Science.
3. Milanesi, M., & Guercini, S. (2023). Image-based Social Media and Visual Content Analysis: Insights from a Literature Review.
4. Ling, C., et al. (2021). Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes.
5. Phan, H. K., et al. (2021). Deep Multimodal Meme Classification for Identifying Hateful Memes, CVPRW.
6. Highfield, T., & Leaver, T. (2016). Instagrammatics and Digital Methods: Studying Visual Social Media, from Selfies and GIFs to Memes and Emoji.
7. Kumar, A., & Garg, G. (2020). Sentiment Analysis of Multimodal Twitter Data.
8. Han, Y. (2022). Design a Meme: Visual Representation, Creative Strategies and Memetic Culture (Thesis).
9. Wang, X., et al. (2023). Visual Sentiment Analysis for Social Media Using Transfer Learning and Multimodal Fusion, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM).
10. Hermida, P. C. Q., & dos Santos, E. M. (2024). Detecting Hate Speech in Memes: A Review.
11. Vázquez-Cano, J. L. (2023). Multimodal Approaches for Emotion Recognition in Internet Memes, Multimedia Tools and Applications, Springer.
12. Pandiani, D. S. M., et al. (2024). Toxic Memes: A Survey of Computational Perspectives on the Detection and Explanation of Meme Toxicities.