Multimodal Deep Learning for Online Meme Classification

Conference Paper · December 2024

DOI: 10.1109/BigData62323.2024.10825758

6 authors, including:

Sebastian Leal-Arenas
University of Pittsburgh
7 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Eftim Zdravevski
Saints Cyril and Methodius University of Skopje
210 PUBLICATIONS **3,858** CITATIONS
SEE PROFILE

Charles Casimiro Cavalcante
Federal University of Ceará
190 PUBLICATIONS **1,095** CITATIONS
SEE PROFILE

Zois Boukouvalas
American University
53 PUBLICATIONS **982** CITATIONS
SEE PROFILE

 DOI: 10.1109/BigData62323.2024.10825758

Multimodal Deep Learning for Online Meme Classification

Stephanie Han
Department of Computer Science
American University
Washington, DC, USA
sh3704a@american.edu

Sebastian Leal-Arenas
Department of Linguistics
University of Pittsburgh
Pittsburgh, PA, USA
sal209@pitt.edu

Eftim Zdravevski
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
eftim.zdravevski@finki.ukim.mk

Charles C. Cavalcante
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Ceará, Brazil
charles.casimiro@ieee.org

Zois Boukouvalas
Department of Mathematics and Statistics
American University
Washington, DC, USA
boukouva@american.edu

Roberto Corizzo
Department of Computer Science
American University
Washington, DC, USA
rcorizzo@american.edu

***Abstract*—Memes possess a humorous intent, yet they can also be used for malicious purposes. Analysing meme data has the potential to enhance content monitoring, identify emerging topics, and support content moderation in online platforms. Memes also represent an interesting use case for multimodal machine learning, as they combine text and image data. In this study, we explored the linguistic characteristics and analysed the convergent themes of five meme classes through common word extraction. Moreover, we compared the effectiveness of various machine learning models, i.e., unimodal (text or image) and multimodal (early fusion, late fusion) in binary and multi-class meme classification tasks. Our results on a large meme dataset showed that memes heavily adhered to current affairs, demonstrated by the high frequency of topical words across meme classes. Regarding model accuracy, early fusion achieved superior accuracy over late fusion in meme classification. Binary models outperformed multi-class classification methods. However, fusion models did not consistently surpass the accuracy of independent text or image-based models.**

***Index Terms*—Memes, multimodal learning, data fusion, neural networks, deep learning**

I. Introduction

Memes have surged in popularity in recent years, with over 180 million posts across various social media platforms [1]. While many memes are created for harmless, humorous intent, they have also been used to produce and spread hate speech and cyberbullying [2], [3], [4]. Unimodal (text or image) and multimodal (text + image) approaches have been employed to detect hate content in memes. Authors in [5] compared the performance of a unimodal approach, BERT, to enhance the efficiency of current text-based models, against a multimodal approach that detects objects through pre-trained models such as InceptionV3, ResNet50, and Xception, integrating both text and image modalities. Results showed that BERT improved the performance of text-based models for detecting hateful content; however, the multimodal model achieved higher accuracy, while maintaining a lightweight structure suitable for practical applications. Improved accuracy has also been attested with other multimodal approaches such as VisualBERT [6], VGG16-BiLSTM [7] and VL-PTM [8]. The impact of images on model accuracy is highly sensitive to the composition of the training dataset [9], suggesting that an integrated approach combining text and image could enhance the classification performance.

Linguistic aspects of memes have been identified, with vernacular language use, alongside slang and neologisms being usually employed [10]. Research has not extensively compared word counts across specific meme categories, nor the intersections of these words, leading to a gap in understanding how these dynamics may influence content moderation practices. Online communication is inherently transient, with linguistic and cultural references that evolve rapidly over time [11]. As a result, meme classification should not focus exclusively on hate detection, as linguistic and visual cues associated with harmful content may shift. Therefore, accurately categorising memes into specific adaptable categories [12] has potential for online content monitoring, detection of emerging trending topics [13], [14], and it can be effective to facilitate content moderation practices.

In this paper, we explore the linguistic characteristics of different memes and compare performance of three distinct approaches to meme classification: text-only, image-only, and a combined text+image approach, leveraging a large real world meme dataset.

II. Method

A. Image Model

VGG-16 [15] is being used as the base of the image pre-trained model in this study. VGG-16 is a CNN (Convolutional Neural Network) model for image recognition which processes image inputs of (224, 224, 3) and consists of 16 layers. 3x3

convolutions and a pooling layer of 2x2 are used to perform feature extraction.

B. Text Model

We adopt BERT [16] (Bidirectional Encoder Representations from Transformers) as our text model. Published in 2018 by Google, BERT learns text representations from both directions (left to right and right to left), so it can get a better sense of the context of each word. BERT is available in different versions such as *base*, *large*, *uncased* and so on, with over a hundred languages languages including English and Chinese.

In our work, we adopt BERT *base* pre-trained[1], whose architecture consists of 12 transformer blocks, 768 hidden layers, and 12 attention heads. BERT requires specifically formatted inputs, for each tokenized input sentence: "Input Id", "Segment Mask", "Attention Mask" and "Label". "Input Id" is a sequence of integers identifying each input token to its index number in the BERT tokenizer vocabulary. "Segment Mask" is a sequence of 1s and 0s used to identify whether the input is one sentence or two sentences long. "Attention Mask" is a sequence of 1s and 0s as well, 1s for all input tokens and 0s for all padding tokens. For example, if *max_length* is set to 128, an input sentence shorter than 128 will be padded with 0 to reach this length. "Label" is presented as a single numeric value indicating the class of a given text, e.g. 0 or 1 for binary classification.

C. Multimodal Fusion Models

In our work, image features are extracted with VGG-16, while text features are extracted with BERT. Our multimodal model branch has two variants: early fusion and late fusion. Early fusion starts at the feature level, which means that image and text features are concatenated into a unique vector. After concatenation, a fully-connected layer is adopted to further process the latent representation, and the model ends with a classification layer of size 2 (binary classification) or 5 (multi-class classification). A graphical representation of the model is shown in Figure 1.

Late fusion fuses the modalities at the level of classification scores extracted by text and image models. It is commonly used when data sources are significantly varied from each other, e.g., differences in sampling rate, dimensionality, and unit of measurement. A graphical representation of the model is shown in Figure 2. To accelerate the training process, transfer learning is implemented in both fusion methods.

III. Data Description

We adopted the ImgFlip-Scraped Memes Caption dataset[2] built scraping ImgFlip.com meme generator images. There are 81 distinct meme classes, collectively comprising over 200,000 memes. We select distinct subsets of classes for the linguistic analysis and the classification task. For the linguistic analysis, five meme classes were chosen (see Figure 3):

[1]https://huggingface.co/google-bert/bert-base-uncased

[2]https://www.kaggle.com/datasets/abhishtagatya/imgflipscraped-memes-caption-dataset

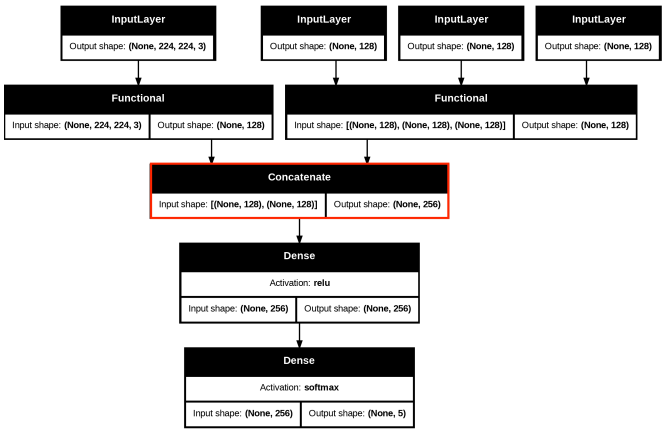

Fig. 1. Early fusion model architecture for multi-class classification.

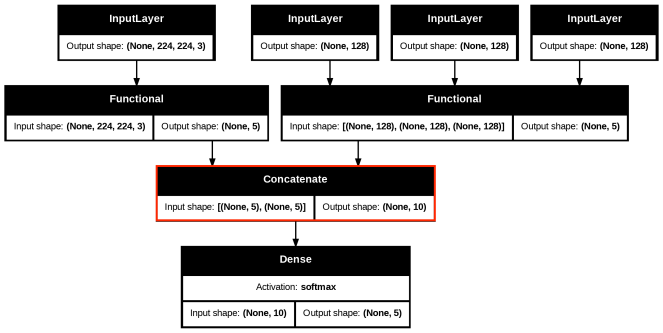

Fig. 2. Late fusion model for multi-class classification.

- *Change My Mind*: It features a photo of the conservative podcaster Steven Crowder sitting at a table with a sign that invites others to change his mind on a controversial statement. It is often used to present a provocative opinion, inviting debate or disagreement.
- *Distracted Boyfriend*: It depicts a man turning his head to look at another woman while his girlfriend looks at him disapprovingly. It is used to illustrate the concept of distraction or shifting attention from one thing to another.
- *Drake Hotline Bling*: It consists of two panels featuring rapper Drake. In the first panel, he is shown rejecting something with a disapproving expression, and in the second, he is smiling and approving of something else. It is commonly used to represent preferences or choices.
- *Laughing Men in Suits*: It shows a group of well-dressed men laughing together, often used to convey a sense of shared amusement or to mock a situation, highlighting hypocrisy or insincerity.
- *Tuxedo Winnie*: It features Winnie the Pooh in two contrasting images – one in a simple, casual form and the other in a tuxedo. It is used to illustrate the idea of duality, often contrasting simplicity with sophistication.

For the binary classification task, we chose **Who Killed Hannibal** and **Scared Cat** classes. For multi-class classification, **Sleeping Shaq**, **Uncle Sam**, and **Peter Parker Cry** were added. Data were split into training (75%) and testing (25%)

Change My Mind | Distracted Boyfriend | Drake Hotline Bling | Laughing Men In Suits | Tuxedo Winnie The Pooh

Fig. 3. Examples of memes in the ImgFlip-Scraped Memes Caption dataset – classes selected for linguistic analysis.

for both binary and multi-class classification tasks. The class distribution ratios in the training and testing sets are almost perfectly balanced for both classification settings: binary (0.52-0.48: training, 0.51-0.49: testing); multi-class (0.23-0.20-0.20-0.19- 0.17: training, 0.25-0.21-0.19-0.19-0.16: testing).

IV. Experiments

Common words extraction: We grouped data based on the meme class label, and counted the frequency of each word. We filtered stopwords as a common practice to remove non-lexical words [17]. We resorted to the pre-curated list available in the Python *nltk* library. We augmented this list with additional stopwords extracted from the top 30 most-common words across all meme classes. From the filtered set of words, we included the 10 most common words for each meme class.

Image model (VGG-16): The activation functions used in the image model are sigmoid for binary classification and softmax for multi-class classification. We used the Adam optimizer with a learning rate of 0.005. While compiling the model, the chosen loss functions are binary cross-entropy and categorical cross-entropy, respectively. The output layer of the models had a size of 1 (binary) or 5 (multi-class). Both models were trained for 10 epochs.

Text model (BERT): The activation used for the text model is sigmoid for binary classification and softmax for multi-class classification. The learning rate was set as default to $3e-5$ using the Adam optimizer. The loss functions used when compiling the models were the same as the image model. We trained two different text models with an output layer of size 1 (binary) and 5 (multi-class).

Multimodal model: After building the image and text-only models, we devised multimodal early and late fusion models [3].

Early fusion: The Image branch of the model had a fully connected output layer of size 128, which matched the Text model's input *max_length*. The fully connected output layer of the Text branch had size 256 due to a Bi-directional LSTM layer added to fine-tune the model. After concatenation, the output layer of the early fusion model with size corresponding to the number of classes subject to prediction (2 or 5). The learning rate used to train the model is set to $3e-5$.

Late Fusion: Instead of fusing 128 and 256 embedding features for image and text as in the early fusion method, this model fused the fully connected output layers of size 5 (multi-class) for both model branches into a new classification layer of size 5, leading to a combined prediction score. All other model specifications are the same settings as the early fusion model.

[3]Our code implementation is publicly available: https://github.com/Stephanie9606/MultimodalDeepLearning_OnlineMemeClassification

TABLE I
Most common words for meme classes selected for linguistic analysis.

Change My Mind	Distracted Boyfriend	Drake Hotline Bling	Laughing Men in Suits	Tuxedo Winnie
coronavirus (124)	fortnite (169)	making (120)	Trump (160)	water (63)
better (85)	minecraft(148)	coronavirus (115)	says (127)	coronavirus (50)
mind (77)	coronavirus (101)	paper (89)	think (107)	using (47)
change (67)	paper (84)	using (89)	president (96)	getting (42)
best (60)	girl (78)	getting (86)	Hillary (79)	19 (41)
toilet (55)	toilet (77)	toilet (84)	us (68)	mother (41)
upvotes (53)	old (77)	minecraft(82)	joke (62)	covid (40)
water (53)	corona (75)	corona (80)	going (59)	eating (40)
paper (51)	homework (72)	fortnite (78)	ha (57)	pooh (40)
made (49)	covid (63)	19 (76)	still (53)	play (34)

TABLE II
Summary of experimental results.

	Binary			Multi-Class		
	Training set		Test set	Training set		Test set
	Epoch 1	Epoch 10	Test	Epoch 1	Epoch 10	
Text	0.86	0.99	0.87	0.82	0.99	0.84
Image	0.93	0.99	1.00	0.32	0.99	0.99
Early fusion	0.56	0.99	1.00	0.41	0.99	0.99
Late Fusion	0.47	0.97	0.99	0.23	0.61	0.61

The following research questions were addressed:

RQ1: What are the most frequently used words across five meme categories, and do these categories exhibit convergent themes?

RQ2: Does the multimodal model outperform the text-only and image-only approaches?

V. Results and Discussion

RQ1: Table I shows the most frequently used words in the selected meme classes. The Change my Mind meme group had a high count for words related to opinion and debate, e.g., 'better,' 'change,' 'upvote,' while the Laughing Men in Suits had a strong political emphasis, frequently featuring terms such as 'Trump,' 'president,' and 'Hillary.' The Distracted Boyfriend group featured the highest counts for gaming terms. The Drake Hotline Bling and the Tuxedo Winnie the Pooh meme groups presented a balanced mixed of references to popular culture and everyday elements. Given the collection date of the dataset, words related to the COVID-19 pandemic, e.g., 'coronavirus,' 'COVID,' 'corona,' appeared in multiple groups. Similarly, 'paper' and 'toilet' were employed in three of the memes, indicating that memes adapted to and commented on contemporary issues. Gaming culture was heavily present in the

memes, with references to 'Fortnight' and 'Minecraft' appearing in three out of the five meme classes.

RQ2: The accuracy of the different models is presented in Table II and Figure 4.

Image model: The binary model achieved an accuracy of 0.93 in epoch 1, while the multi-class model had only 0.32. At epoch 10, both models had an accuracy of 0.99 (training set).

Text model: Our results showed a similar accuracy for both models: In the first epoch both models had an accuracy of more than 0.8 (binary: 0.86, multi-class: 0.82). At the end of the training stage, both models reached an accuracy of 0.99.

Fusion models: In binary classification, early and late fusion did not present notable differences. These models presented accuracy scores close to 0.5 in the first epoch, both reaching 0.99 in epoch 10. The possible reason for this outcome is that there were only two classes, which is a significantly easier task than multi-class classification.

Regarding multi-class classification, results were significantly different. The early fusion model achieved an accuracy of just 0.41 in the first epoch, and eventually reached 0.99 at the end of the training stage, as in other models. In contrast, the late fusion model achieved very low accuracy of 0.23 in the first epoch, and 8 epochs were necessary to reach an accuracy of over 0.5. It achieved an accuracy of 0.61 at the end of the training process, which is a sub-par result compared to early fusion. One possible explanation for this phenomenon is a limited flexibility for the late fusion model, which operated on single model inferences instead of fusion at a feature level. On the other hand, early fusion is more effective at modeling signal-level interactions between modalities.

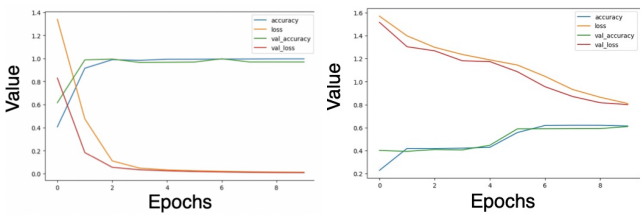

Fig. 4. Left: Early fusion result of the multi-class classification model; Right: Late fusion result of the multi-class classification model

VI. Conclusion and Future Work

In this paper, we studied the linguistic characteristics of different memes and compared the performance of three distinct approaches to meme classification: text-only, image-only, and text+image. Our results showed that popular culture and pandemic-related terms were common in the dataset, with divergence focus and tone. Some memes primarily engaged with personal decision-making and humour, whereas others adopted a more critical, political stance. The presence of common terms also highlighted the adaptability of memes in reflecting societal concerns and cultural references.

A comparison of multimodal approaches for meme classification showed a higher accuracy for early fusion as opposed to late fusion. Binary models outperformed multi-class models in accuracy, due to the simpler classification task with fewer label categories. Fusion models did not consistently surpass independent image or text models in accuracy given the similarity of images in the dataset or the limitation of training with 5 classes rather than the full set of 81 classes.

Future research could expand the classification task to include the full set of classes, potentially improving model accuracy and generalisability. Investigating hybrid approaches that combine early and late fusion may also yield insights into more effective ways of integrating text and image features. Incorporating a larger and more diverse dataset with updated cultural references could help to assess the model's adaptability and relevance across different contexts and time frames.

References

[1] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil, "On the origins of memes by means of fringe web communities," in *Proceedings of the internet measurement conference 2018*, 2018, pp. 188–202.

[2] P. Aggarwal, M. E. Liman, D. Gold, and T. Zesch, "Vl-bert+: Detecting protected groups in hateful multimodal memes," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 2021.

[3] P. C. d. Q. Hermida and E. M. d. Santos, "Detecting hate speech in memes: a review," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12 833–12 851, 2023.

[4] J. Badour and J. A. Brown, "Hateful memes classification using machine learning," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021.

[5] A. Bhat, V. Varshney, V. Bajlotra, and V. Gupta, "Detection of hatefulness in memes using unimodal and multimodal techniques," in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2022, pp. 65–73.

[6] R. Velioglu and J. Rose, "Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge," *arXiv preprint arXiv:2012.12975*, 2020.

[7] M. T. Ahmed, N. Akter, M. Rahman, D. Das, and R. G. AZM T, "Multimodal cyberbullying meme detection from social media using deep learning approach," *Int J Comput Sci Inf Technol (IJCSIT)*, vol. 15, pp. 27–37, 2023.

[8] Y. Chen and F. Pan, "Multimodal detection of hateful memes by applying a vision-language pre-training model," *Plos one*, vol. 17, no. 9, p. e0274300, 2022.

[9] P. Aggarwal, J. Mehrabanian, W. Huang, Ö. Alaçam, and T. Zesch, "Text or image? what is more important in cross-domain generalization capabilities of hate meme detection models?" *arXiv preprint arXiv:2402.04967*, 2024.

[10] B. Kostadinovska-Stojchevska and E. Shalevska, "Internet memes and their socio-linguistic features," *European journal of literature, language and linguistics studies*, vol. 2, no. 4, 2018.

[11] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 177–186.

[12] D. Kotsakos, P. Sakkos, I. Katakis, and D. Gunopulos, "Language agnostic meme-filtering for hashtag-based social network analysis," *Social Network Analysis and Mining*, vol. 5, pp. 1–14, 2015.

[13] F. Abousaleh, W. Cheng, N. Yu, and Y. Tsao, "Multimodal deep learning framework for image popularity prediction on social media," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[14] C.-Y. Chiu, H.-Y. Lane, J.-L. Koh, and A.-L. Chen, "Multimodal depression detection on instagram considering time interval of posts," *Journal of Intelligent Information Systems*, vol. 56, no. 1, pp. 25–47, 2020.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[17] R. Corizzo and S. Leal-Arenas, "One-class learning for ai-generated essay detection," *Applied Sciences*, vol. 13, no. 13, p. 7901, 2023.