

Classifying the Relation Between Images and Text on Social Media Using Vision-Language Models

Edison Marrese-Taylor Matīss Rikters
National Institute of Advanced Industrial Science and Technology
{firstname.lastname}@aist.go.jp

Abstract

Social media websites have had the option of multimedia uploads for more than a decade now. However, the relation between the text and the posted images is not always unambiguous if there is a relation at all. We explore how multilingual vision-language models tackle the task of image-text relation prediction in different languages, and prepare dedicated balanced benchmark data sets from Twitter posts in Latvian and English. We compare our results to previous work and show that the more recently released vision-language model checkpoints are becoming increasingly capable at this task, but there is still much room for further improvement. Experiments with in-context learning outline how further improvements can be achieved.

1 Introduction

The social network formerly known as Twitter (now X¹⁾) remains a crucial platform in modern society due to its role in shaping public discourse, enabling real-time communication, and fostering global conversations. As a microblogging site, it allows individuals, organizations, and governments to share thoughts, news, and opinions instantaneously. Even though potential alternatives have risen in popularity in the past years, they still exhibit distinct drawbacks to the general public. For example, Threads is still refusing to promote real-time content and news events, or Mastodon being too granulated and slow overall due to being dependent on the performance of individual servers.

In 2011 when Twitter integrated posting images along with text, it enhanced the platform's impact by offering a visual dimension to amplify the message. Images can serve as powerful tools to evoke emotional responses, clarify

1) From Twitter to X: Elon Musk Begins Erasing an Iconic Internet Brand - <https://www.nytimes.com/2023/07/24/technology/twitter-x-elon-musk.html>



Text	Es nezinu, kādu narkotiku cepēji pievieno šim gardumam, bet es to varētu ēst un ēst bez apstājas līdz pat pavasarim.
Translation	I don't know what kind of drug the bakers add to this treat, but I could eat it and eat it non-stop until spring.

Figure 1 An example of an attached image to a tweet which is difficult to comprehend without the text, as well as the text cannot be fully explained without seeing the image.

complex issues, and influence perceptions, but that is not always the case. The images can also be added just as an attention-grabbing strategy or clickbait, or even expressing humor as a meme. A tweet accompanied by a striking or controversial image can dramatically shift how readers interpret the message, adding layers of meaning or even altering the context. In this way, the synergy between text and visuals on the social network not only grabs attention but also guides the overall narrative.

In this work, we extend previous research by Vempala and Preoțiuc-Pietro [1] and Rikters et al. [2] who introduced a four-class taxonomy for classifying image-text relations from Twitter data and performed initial experi-

ments with early versions of LLaVA [3] models. We further divide the test set published by Rikters et al. [2] into a class-balanced evaluation set to lessen the overarching dominance of specific classes. We also employ a professional translator to manually translate their evaluation set from Latvian into English to minimize the potential errors that could be introduced by using automatic translations for the vision-language model (VLM) experiments. We experiment with four different open-source VLM checkpoints that are capable of running on consumer hardware.

2 Related Work

Vempala and Preoțiu-Pietro [1] introduced the categorization schema for the relations between Tweet text and attached images that we use in our experiments. They distinguished four different categories: 1) the image adds to the text meaning and the text is represented in the image (further in the paper we will denote this using the emoji combination 🗃️✅📄✅); 2) the image adds to the text meaning and the text is not represented in the image (🗃️✅📄❌); 3) the image does not add to the text meaning and the text is represented in the image (🗃️❌📄✅); and 4) the image does not add to the text meaning and the text is not represented in the image (🗃️❌📄❌). They also released a 4472 tweet-image pair corpus with manually annotated relation categories (2942 were available at the time of writing this paper) and analyzed the user demographic traits linked to each of the four image tweeting categories in depth.

Rikters et al. [2] applied the image-tweet categorization schema introduced by Winata et al. [1] on the Latvian Twitter Eater Corpus (LTEC) by annotating 812 tweets written in Latvian about topics related to food and eating. They experimented with automatically classifying the original data set of Latvian tweets, as well as automatic translations of the texts into English, using LLaVA models of versions 1.3 and 1.5 in sizes of 7B and 13B parameters. They reported results of 20.69% prediction accuracy when evaluated on the original Latvian texts, and increasing up to 27.83% when evaluated on the automatic English translations.

Winata et al. [4] released a massively multilingual data set of food-related text-image pairs for visual question answering by identifying dish names and their origins in 30 languages. They evaluated these tasks using various VLMs in multiple sizes and release open-source code for experiment reproduction. Their results showed that closed

proprietary online API systems show overall superior performance, however, open-source models in the 70B-90B parameter range can still be quite competitive.

3 Proposed Approach

We commit to a more detailed evaluation of the image-text relation classification task for the available Twitter data. We aim to compare the performance of several recent VLMs that can be run on a reasonable desktop setup using a single NVIDIA RTX 3090 GPU with 24GB of VRAM. In our evaluation will consider the following model versions and sizes – LLaVa-NeXT Vicuna [5] 7B and 13B, Qwen2-VL [6] 7B, Phi 3.5 Vision [7] 4B, which we load from the Hugging Face model repository.

Our evaluation is based on the LTEC image-text relation test set in Latvian and manual translations of the texts into English. The test set is reduced in size to favor a more balanced class distribution, enabling a fair evaluation. In addition to the overall class, we also present a separate evaluation of the two individual questions prompted to the models - Q1) is the image adding to the text meaning; and Q2) is the text represented to the image.

We also attempt to improve the results by using in-context learning [8] by providing several examples of the image-text relation task at each inference step, and consider the applicability of further fine-tuning VLMs on the image-text relation task.

4 Data Preparation

The previous work which evaluated the image-text relations using VLMs exhibited several flaws. Firstly, the data set composition was skewed strongly towards two of the four classes as shown in Table 2 - the image adding to the text meaning and text being represented in the image class with 48.28% of the data and a further 36.45% for the image not adding to the text meaning and text being represented in the image class, which together make up 84.73% of the evaluation data. Furthermore, the authors did not report separate evaluation on the individual question performance that were prompted to the VLMs. Finally, the evaluation which achieved the highest accuracy result was performed on automatically translated texts, which could be erroneous, making way for the potential of creating further unnecessary errors in the classification task.

System	BLEU	ChrF	COMET
Tilde MT	52.63	67.94	78.50
Google Translate	63.49	75.56	83.99
DeepL Translate	59.19	72.20	83.31
Opus MT	54.50	68.77	78.78

Table 1 Machine translation results





















Class	Tweets	Percentage	Before
   	113	32.29%	48.28%
   	72	20.57%	8.87%
   	113	32.29%	36.45%
   	52	14.86%	6.40%

Table 2 Evaluation set class distribution.  represents the image adding to the text meaning,  – the text being represented in the image, and  and  – true or false respectively.

4.1 Manual Translation

The highest text-image relation classification accuracy scores reported by Rikters et al. [2] were achieved by automatically translating the Latvian texts into English using an MT system that reaches scores of 48.28 BLEU and 68.21 ChrF on a separate evaluation score. While MT systems of such quality are generally usable, they are still far from perfect. To minimize the potential of error propagation we employed a human translator to perform a full manual translation of the image-tweet relation texts from Latvian into English. We evaluated three online systems²⁾ and one open-source model³⁾ on the manually translated texts. Results in Table 1 show that for this set Google Translate seems to be outperforming all others according to BLEU [9], ChrF [10] and COMET[11], and Tilde MT, which was used in the evaluation of Rikters et al. [2] scores the lowest. In the subsequent evaluations of this paper, we only use our manual translations of the Latvian tweets when referring to the English translations.

4.2 Evaluation Set Balancing

We divided the 812 tweet set into a separate evaluation set of 350 tweets to have a more even distribution among the four classes. The main objective was to reduce the dominance of the first and third classes. A comparison of the new distribution with the full original data set is shown in Table 2.

2) Tilde MT, Google Translate, DeepL Translate - all accessed in November 2024

3) Opus MT tc-big-lv-en: <https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-lv-en>

4.3 Instruction Formatting

It is well known that many modern large language models and therefore also VLMs can often be very sensitive to the provided prompt for a specific task and produce vastly variable results. In our experiments, we mainly kept using the prompt suggested by Rikters et al. [2] for all models.

5 Results

Our main results are summarized in Table 3. We compare four different models which represent 3 main size categories of 4B, 7B and 13B parameters. Each evaluation is run with 10 different seeds (the same seeds for each model) with the prompt written in English and the actual tweet text provided in either Latvian or English. We compare classification accuracy on the overall class, as well as each of the two individual questions of the image adding to the meaning and text being represented in the image.

The result table shows great variation in both the overall class accuracy, as well as the individual questions. Best results are achieved by the LLaVA-NeXT models and Phi 3.5, of which all seem to prefer the English translation rather than the original Latvian text. Qwen2-VL scores the lowest, regardless of the input language, and also exhibits no variation with the different random seeds. Meanwhile, Phi 3.5 and especially LLaVA-NeXT models tend to vary a lot. The LLaVA-NeXT outperform the results reported by Rikters et al. [2], although they are not directly comparable.

As an ablation study, we also experimented with providing both the text and the instruction prompt in Latvian, however, this led to mostly very incomplete results. Table 4 summarizes our findings where out of the 350 tweets many were not answered directly with yes/no as requested in the prompt required manual verification of the model output whether it contains the answer at all. In addition, up to 21% of the model outputs did not contain an answer in the case of LLaVA-NeXT 13B.

For comparison, we also sampled a random subset of 450 tweets from the larger data set by Winata et al. [1] for evaluation. This data set seems to be naturally better distributed already, having a distribution of 19.33% : 24.89% : 23.33% : 32.45%. Accuracy results in Table 5 do show overall higher scores than the very domain-specific Latvian food tweets, but in general they are still relatively low and have the potential to be further improved.

Prompt	Data	Model	Class	Question 1	Question 2
EN	LV	LLaVA-NeXT 7B	23.40 ± 8.03	51.57 ± 3.57	41.37 ± 21.49
EN	LV	LLaVA-NeXT 13B	19.43 ± 4.57	51.11 ± 6.03	34.60 ± 3.11
EN	LV	Phi 3.5 4B	18.14 ± 3.00	48.49 ± 1.63	38.71 ± 3.57
EN	LV	Qwen2-VL 7B	15.71 ± 0.00	47.71 ± 0.00	35.43 ± 0.00
EN	EN	LLaVA-NeXT 7B	24.46 ± 7.83	52.17 ± 1.31	43.86 ± 18.71
EN	EN	LLaVA-NeXT 13B	28.91 ± 6.34	53.20 ± 4.06	51.40 ± 10.89
EN	EN	Phi 3.5 4B	25.14 ± 5.71	48.31 ± 2.83	49.14 ± 7.43
EN	EN	Qwen2-VL 7B	15.71 ± 0.00	47.43 ± 0.00	37.14 ± 0.00

Table 3 Average classification accuracy results from zero-shot experiments using 10 different random seeds.

Model	Class	Q1	Q2	Remarks
LLaVA 7B	32.43 ± 0.53	52.80 ± 0.63	64.54 ± 0.60	1-3 cases were not answered
LLaVA 13B	28.41 ± 5.99	43.46 ± 3.69	45.89 ± 4.17	58-75 cases not answered
Qwen2-VL	15.43 ± 0.00	47.43 ± 0.00	37.43 ± 0.00	
Phi 3.5	19.92 ± 8.31	50.54 ± 3.69	37.17 ± 12.26	5-17 cases were not answered

Table 4 Results from prompting tweets in Latvian with a Latvian prompt.

	Class	Q1	Q2
LLaVA 7B	32.29±13.27	48.04±5.51	64.51±20.16
LLaVA 13B	36.82±5.40	55.53±5.80	64.64±4.64
Qwen2-VL	33.11±0.00	55.56±0.00	59.11±0.00
Phi 3.5	35.47±4.98	62.22±2.67	57.98±3.13

Table 5 Results from a random subset of 450 English Tweets from Vempala and Preotjiuc-Pietro [1].

ICL	LLaVA-NeXT		Phi 3.5 4B	Qwen 2 7B
	7B	13B		
0	24.46±7.83	28.91±6.34	25.14±5.71	15.71
1	25.97±8.54	25.63±7.34	19.51±2.37	17.43
2	27.37±9.09	24.77±5.91	25.37±8.63	18.29
3	26.77±8.49	23.83±5.89	21.51±9.34	15.71
4	28.17±9.89	23.66±6.34	20.11±5.31	15.71
5	29.11±8.83	23.09±7.77	19.89±5.03	16.00

Table 6 In-context learning experiment class-wise classification accuracy without providing an image.

5.1 In-context learning

To further improve the results, we experiment with using in-context learning (ICL) [8] by providing several examples of the image-text relation task at each inference step. While all of our chosen models do support multi-image inference, we experienced very unstable performance when evaluating, therefore we chose to run ICL experiments using only text for the in-context examples. We experimented by providing the models with 1 to 5 sets of examples where each set includes one example of each of the four classes.

Results in Table 6 show varied success with the best results from the zero-shot experiments – LLaVA-NeXT 13B and Phi 3.5 4B – not improving at all. However, LLaVA-NeXT 7B was able to gradually improve with each additional set of ICL examples, and Qwen 2 7B also demonstrated slightly increased performance with 2 ICL examples.

6 Conclusion

In this paper, we introduced an extended evaluation of the image-text relation task for social media posts from

Twitter. We prepare balanced versions of previously available image-text relation data sets, as well as a manual English translation of the original Latvian texts. We experiment with several open-source vision-language models and demonstrate how results vary depending on multiple conditions. Initial experiments with in-context learning highlight the potential applicability of this method for further improvements.

We plan to release our balanced evaluation data set along with the code that we used for our evaluation for easy reproduction of our results or similar experiments. In future work we intend to explore further applicability of the in-context learning approach, as well as perform fine-tuning on the model checkpoints for the image classification task.

Acknowledgements

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

- [1] Alakananda Vempala and Daniel Preoțiuc-Pietro. Categorizing and inferring the relationship between the text and image of Twitter posts. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 2830–2840, Florence, Italy, July 2019. Association for Computational Linguistics.
- [2] Matīss Rikters, Rinalds Vīksna, and Edison Marrese-Taylor. Annotations for exploring food tweets from multiple aspects. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 1233–1238, Torino, Italia, May 2024. ELRA and ICCL.
- [3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.
- [4] Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohelj, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Santosa, Peerat Limkonchotiawat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie Lee, Shogo Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines, 2024.
- [5] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024.
- [6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.
- [7] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Ye-long Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.
- [8] Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. VI-icl bench: The devil in the details of multimodal in-context learning, 2024.
- [9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [10] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In **Proceedings of the Tenth Workshop on Statistical Machine Translation**, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
- [11] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.