RAEmoLLM: Retrieval Augmented LLMs for Cross-Domain Misinformation Detection Using In-Context Learning Based on Emotional Information

Zhiwei Liu[1] **Kailai Yang**[1] **Qianqian Xie**[2] **Christine de Kock**[3]
Sophia Ananiadou[1*] **Eduard Hovy**[3,4]
[1] The University of Manchester [2] School of Artificial Intelligence, Wuhan University
[3] The University of Melbourne [4] Carnegie Mellon University
{zhiwei.liu,kailai.yang,sophia.ananiadou}@manchester.ac.uk
xqq.sincere@gmail.com, {christine.dekock,eduard.hovy}@unimelb.edu.au

Abstract

Misinformation is prevalent in various fields such as education, politics, health, etc., causing significant harm to society. However, current methods for cross-domain misinformation detection rely on effort- and resource-intensive fine-tuning and complex model structures. With the outstanding performance of LLMs, many studies have employed them for misinformation detection. Unfortunately, they focus on in-domain tasks and do not incorporate significant sentiment and emotion features (which we jointly call *affect*). In this paper, we propose RAEmoLLM, the first retrieval augmented (RAG) LLMs framework to address cross-domain misinformation detection using in-context learning based on affective information. RAEmoLLM includes three modules. (1) In the index construction module, we apply an emotional LLM to obtain affective embeddings from all domains to construct a retrieval database. (2) The retrieval module uses the database to recommend top K examples (text-label pairs) from source domain data for target domain contents. (3) These examples are adopted as few-shot demonstrations for the inference module to process the target domain content. The RAEmoLLM can effectively enhance the general performance of LLMs in cross-domain misinformation detection tasks through affect-based retrieval, without fine-tuning. We evaluate our framework on three misinformation benchmarks. Results show that RAEmoLLM achieves significant improvements compared to the other few-shot methods on three datasets, with the highest increases of 15.64%, 31.18%, and 15.73% respectively. This project is available at https://github.com/lzw108/RAEmoLLM.

1 Introduction

The internet is flooded with misinformation (Scheufele and Krause, 2019), which has a significant impact on people's lives and societal stability (Della Giustina, 2023). Misinformation is pervasive across various domains such as education, health, technology, and especially on the internet, which requires people to invest significant time and effort in discerning the truth (Pérez-Rosas et al., 2018). However, models trained in specific known domains are often fragile and prone to making incorrect predictions when presented with samples from new domains (Saikh et al., 2020). As a result, detecting cross-domain misinformation has become an urgent global issue and poses greater challenges and difficulties.

Although some studies address cross-domain misinformation detection (Comito et al., 2023; Tang et al., 2023; Shi et al., 2023), they require effort-intensive fine-tuning, and apply only traditional machine learning methods or complex deep learning methods. Recently, LLMs have achieved impressive results in various tasks through zero-shot, few-shot (Li, 2023), or instruction tuning (Zhang et al., 2023a). Many researchers have applied LLMs to identify misinformation (Li et al., 2023; Hu et al., 2024; Cheung and Lam, 2023). However, these methods perform only in-domain misinformation detection. Moreover, emotions and sentiments (which we jointly call *affect*) are important characteristics of human expression and communication (Hakak et al., 2017). When authors publish misinformation, they often consciously choose specific emotions to capture the attention and resonance of readers to encourage rapid spread (Keen, 2006; Liu et al., 2024d). Unfortunately, there are few LLMs that utilize affective information to detect misinformation, and the only ConspEmoLLM (Liu et al., 2024b) are developed based on an emotional LLM, which does not make full use of affective information, has no cross-domain ability, and also needs time-consuming fine-tuning.

In-context learning (ICL) needs only task instructions and few-shot examples (input-label pairs),

*Corresponding Author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16508–16523
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

eliminating fine-tuning on specific task labels (Dong et al., 2022b). A few studies have used ICL to address cross-domain problems (Long et al., 2023; Wu et al., 2024). To the best of our knowledge, there is currently no application of ICL for cross-domain misinformation detection based on affective information retrieval.

To address these issues, we propose the first retrieval augmented (RAG) LLMs framework based on emotional information (RAEmoLLM), to address cross-domain misinformation detection using in-context learning based on affective information. RAEmoLLM contains three modules: (1) In the *index construction* module, we apply EmoLLaMA-chat-7B (Liu et al., 2024c) to encode all domain corpora, obtaining implicit affective embeddings to construct the retrieval database as well as explicit affective labels. We also conduct a comprehensive affective analysis to demonstrate the effectiveness of affective information for discriminating between true and misinformation. (2) The *retrieval* module recommends the top K affect-related examples (text-label pairs) from the source domain corpus according to the target domain content, obtained from the retrieval database. (3) These examples are utilized as the few-shot demonstrations in the *inference* module, which is driven by a prompt template to guide the LLM to verify the target content for misinformation. The template helps combine implicit and explicit affective information. This framework effectively enhances the capabilities of LLMs in multiple cross-domain misinformation detection tasks through leveraging affective information, without the need for fine-tuning.
In this work, we make three main contributions:

- We conduct affective analysis on different kinds of misinformation datasets and construct the retrieval database according to the implicit affective information for misinformation datasets.

- We propose RAEmoLLM, the first framework for cross-domain misinformation detection using ICL based on affective information, which does not require fine-tuning. Experimental results show that RAEmoLLM outperforms the zero-shot method and other few-shot methods.

- We evaluate RAEmoLLM on a variety of misinformation benchmarks, including fake news, rumours, and conspiracy theory datasets. Results show that RAEmoLLM achieves significant improvements compared to the other few-shot methods on three datasets, with the highest increases of 15.64%, 31.18%, and 15.73% respectively, which illustrate the effectiveness of RAEmoLLM framework.

2 Methodology

This section introduces our method of cross-domain misinformation detection, using the *index construction* module, *retrieval* module and *inference* module. The overall architecture of RAEmoLLM is shown in Figure 1. In the *index construction* module (Sec. 2.1), we collect domain datasets, and employ an emotional LLM to obtain affective embeddings as well as affective labels to conduct a comprehensive affective analysis on them to detect the affective differences between real and false information. The implicit embeddings are adopted to construct the retrieval database, which will be used by the *retrieval* module (Sec. 2.2) to obtain source-domain examples. These results are used as the few-shot examples for *inference* module's (Sec. 2.3) in-context learning to detect target domain misinformation.

2.1 Index Construction Module

In this section, we first introduce the original datasets and the processing procedure at Sec. 2.1.1. We subsequently conduct affective analysis on these datasets and present how and why to obtain implicit and explicit affective information at Sec. 2.1.2. Finally, we apply the implicit affective information to construct the retrieval database (Sec. 2.1.3).

2.1.1 Datasets

We collect FakeNewsAMT (Pérez-Rosas et al., 2018), Celebrity (Pérez-Rosas et al., 2018), PHEME (Kochkina et al., 2018), and COCO (Langguth et al., 2023) datasets. The statistics of these datasets are presented in Table 1. FakeNewsAMT is a cross-domain dataset, including six domains. The legitimate news in FakeNewsAMT was obtained from various mainstream news websites. The authors adopted crowdsourcing via Amazon Mechanical Turk (AMT) to generate fake versions of legitimate news items. The Celebrity dataset was derived from online magazines. We combine FakeNewsAMT and Celebrity as AMTCele. PHEME contains a collection of Twitter rumours and non-rumours posted during nine breaking news events. COCO dataset consists

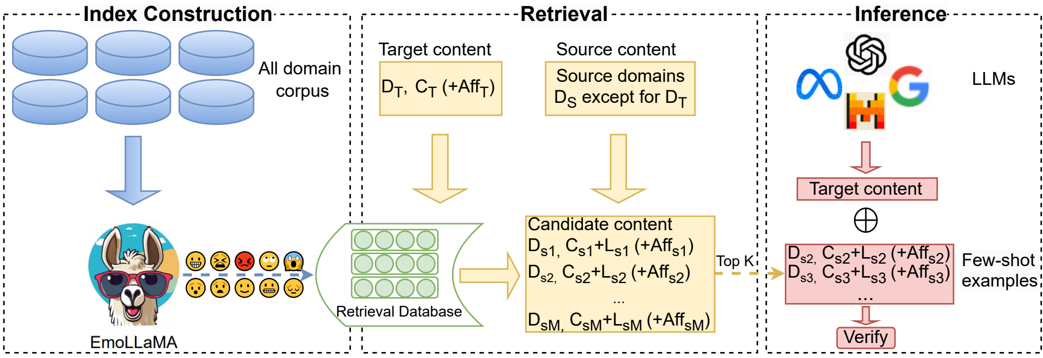

Figure 1: The architecture of RAEmoLLM. D: Domain. T: Target domain. S: Source domain. C: Corpus. L: Label. Aff: Affective information. M: Number of source domain data. Index Construction Module: Apply an emotional LLM to obtain affective embeddings to construct a retrieval database. Retrieval Module: Recommend top K examples (text-label pairs) from source domain data. Inference Module: Adopt the recommended examples as demonstrations for inference.

AMTCele			PHEME			COCO		
Domain	Legit	Fake	Events	Rumours	Non-rumours	Topics	Related	Conspiracy
Technology	40	40	Charlie Hebdo	458	1621	Fake Virus		
Education	40	40	Sydney siege	522	699	Harmful Radiation	248	612
Business	40	40	Ferguson	284	859	Depopulation		
Sports	40	40	Ottawa shooting	470	420	Other 9 domains	540	1181
Politics	40	40	Germanwings-crash	238	231	Total	788	1793
Entertainment	40	40	Putin missing	126	112			
Celebrities	250	250	Prince Toronto	229	4			
Total	490	490	Gurlitt	61	77			
			Ebola Essien	14	0			
			Total	2402	4023			

Table 1: Statistics of datasets. AMTCele includes 7 domains. PHEME contains 9 domains (events). COCO has 12 domains (topics). For AMTCele and PHEME, we apply leave-one-domain-out strategy for evaluation. For COCO, we select 3 domains as test set.

Datasets	Affective	sub-emotion	legit/non-rumours/related		fake/rumours/conspiracy		t-test	
			mean	var	mean	var	t	p
AMTCele	EIreg	Anger	0.3584	0.0064	0.4055	0.0060	-9.3294	6.91E-20
		Fear	0.3587	0.0137	0.4047	0.0124	-6.2861	4.90E-10
		Joy	0.3392	0.0180	0.2897	0.0142	6.1054	1.48E-09
		Sadness	0.3341	0.0109	0.3697	0.0106	-5.3726	9.70E-08
	Vreg	-	0.5471	0.0204	0.4940	0.0170	6.0656	1.88E-09
PHEME	EIreg	Sadness	0.5215	0.0152	0.5177	0.0182	1.1442	0.2526
COCO	Vreg	-	0.3961	0.0095	0.3973	0.0066	-0.3325	0.7395

Table 2: Statistics values of EIreg and Vreg on different datasets. The t-test is conducted between *legit/non-rumours/related* and *fake/rumours/conspiracy*. The complete statistics on PHEME and COCO can be found in Table 13 in the Appendix G.

of 12 conspiracy theory categories[1]. Each tweet in COCO is assigned an overall intention label, as follows: *Conspiracy* is assigned to tweets for which the tweet is related to at least one of the 12 categories and is actively spreading conspiracy theories. Otherwise, if the tweet is related to the specific category, but it does not propagate misinformation or conspiracy theories, then the overall label of *Related* is used. The overall label of *Unrelated* is only used for tweets that are unrelated to all 12 conspiracy categories. We remove the *Unrelated* text since the aim of the cross-domain test.

For AMTCele and PHEME, we apply leave-one-domain-out strategy[2] to evaluate the model. For COCO dataset, due to one text data may involve

[1] Suppressed Cures, Behavior Control, Anti Vaccination, Fake Virus, Intentional Pandemic, Harmful Radiation, Depopulation, New World Order, Esoteric Misinformation, Satanism, Other Conspiracy Theory, Other Misinformation.

[2] By sequentially selecting a specific domain as the test set and the remaining domains as the training set, we can evaluate the model's performance on each individual domain and combine these results to obtain a comprehensive assessment of the overall dataset.

one or multiple topics, we select all data points involving the *Fake Virus*, *Harmful Radiation*, and *Depopulation* topics as the test set, and the other topics as the retrieval dataset.

2.1.2 Affective Analysis

We firstly conduct a comprehensive affective analysis after collecting datasets. EmoLLaMA-chat-7B, which has the best overall performance among the EmoLLMs (Liu et al., 2024c), is used for affective analysis. EmoLLaMA-chat-7B can be used to extract five kinds of affective dimensions (which we jointly call affect), including Emotion intensity (EIreg), Emotion intensity classification (EIoc), Sentiment (valence) strength (Vreg), Sentiment (valence) classification (Voc) and Emotion detection (Ec). The detailed introduction can be found in Appendix G.1.

Obtain implicit and explicit affective information: Following the guidelines of EmoLLMs (Liu et al., 2024c), we add prompts provided by EmoLLMs for each data point in order to obtain vectors from the last hidden layer (i.e., 4096d) for each affective dimension, as well as final labels using EmoLLaMA-chat-7B. We subsequently determine the distribution of affective information in different categories in each dataset.

Explicit affective analysis: Table 2 and Table 13 show regression information (i.e., EIreg and Vreg) of final labels. We use the t-test[3] to measure the difference in emotional intensity between two sets of data. The t-value and p-value calculated between *legit/non-rumours/related* and *fake/rumours/conspiracy* demonstrate that there are statistically significant affective differences between the different categories. Figure 3 to Figure 8 and the chi-squared test in Appendix G.2 confirm that other classifications using affective information are also related to misinformation. However, Table 2 also presents some special cases that cannot effectively distinguish real and false information (e.g. EIreg-sadness in PHEME, Vreg in COCO). Liu et al. (2024b) also conducted some experiments that demonstrated that simply utilizing explicit affective information does not enhance the model's capability. Therefore, we introduce implicit affective information.

[3] t-test is a statistical method used to compare whether the difference between the means of two sets of data is significant. It generates a t-value, which is then compared to a t-distribution to determine if the observed difference is significant.

Implicit affective analysis: Table 14 shows statistics of different affective embeddings (i.e. last hidden layer of EmoLLaMA-chat-7B). We perform t-tests on the top-K cosine similarity within categories and across categories. The results indicate that the similarity within categories is significantly higher than across categories, confirming that similar top-K data points are likely to belong to the same category (further analysis can be found in Appendix G.2). We also visualize the data distribution reduced to 3 dimensions using PCA in Figures 9 and 10 in Appendix. It can be observed that different categories are clearly separated in the latent space. All the above demonstrate the close relationship between affective information and misinformation.

2.1.3 Retrieval Database Construction

After obtaining the implicit affective embeddings in the previous step, we proceed to construct a comprehensive retrieval database. This database consists of vectors that encapsulate rich affective information, enabling efficient retrieval and analysis.

2.2 Retrieval Module

Algorithm 1 Retrieval process

Require: Target domain corpus D_T, source domain corpus D_S, retrieval database R.
Ensure: Target domain corpus with top K retrieval examples D_{retri}.

```
E_T ← R(D_T)
E_S ← R(D_S)
for e_t in E_T do
    for e_s in E_S do
        score = cosine(e_t, e_s)
        Sco ← score
    end for
    D_retri ← select top k examples in R(D_S) according to Sco
end for
```

The retrieval database constructed in Sec 2.1 is represented as R. Algorithm 1 shows the retrieval process. In this module, we first process the multi-domain datasets into text-label pairs to obtain the target domain data $D_T = [\{c_{t1}, l_{t1}\}, \{c_{t2}, l_{t2}\}, ..., \{c_{tN}, l_{tN}\}]$ and source domain data $D_S = [\{c_{s1}, l_{s1}\}, \{c_{s2}, l_{s2}\}, ..., \{c_{sM}, l_{sM}\}]$ (c denotes corpus text, and l is the label. N and M are the numbers of target domain data and source domain data respectively). Following that, we obtain the target domain affective embedding $E_T = [e_{t1}, e_{t2}, ..., e_{tN}]$ and source domain affective embedding $E_S = [e_{s1}, e_{s2}, ..., e_{sM}]$ through the embedding retrieval database R based on the corpus texts in D_T and D_S. Subsequently, we traverse the target domain embedding (e_t) in E_T and calculate the similarity values with each source domain embedding e_s from E_S using the cosine method. Finally, we select the top k examples from source domain for each target domain data based on *Sco* to D_{retri}, which will be the few-shot examples for LLM inference.

2.3 Inference Module

We apply template 1 to construct the instruction datasets for inference once we get the top examples for each target domain data. *[task prompt]* denotes the instruction for the task (The different *[task prompts]* for each datasets can be found in Appendix B). *[input text]* is a data item from the target domain data. *[examples]* are the retrieval examples from source domain data (i.e. D_{retri}) and the *[output]* is the output from LLM.

Template 1
Task: *[task prompt]*
Target text: *[input text]*
Here are a few examples: *[examples]*
According to the above information, the label of target text: *[output]*

We also apply template 2 to add explicit affective information. *[affective information]* contains five dimensions described in Section 2.1.2. The format of *[examples]* is "*Text: [text]. [Affective info]: [value]. The label of text: [label]*". Table 6 shows one complete example.

Template 2
Task: *[task prompt]*
Target text: *[input text] + [affective info]*
Here are a few examples retrieved by *[affective info]*: *[examples]*
According to the above information, the label of target text: *[output]*

3 Experiments

3.1 Base Models

- **LLMs:** We apply ChatGPT (gpt-3.5-turbo-0125), GPT-4o[4], Llama3-8b-Instruct, Llama3.2-(1b,3b)-Instruct[5], Gemma-instruct-(2b, 7b) (Team et al., 2024), Mistral-7b-Instruct (Jiang et al., 2023) and Vicuna-(7b, 13b, 33b) (Chiang et al., 2023) as base models to test our methods.
- **PLMs:** We select BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) as fine-tuning baselines. Specifically, one domain is selected as the target domain, other domains are used as the training dataset to fine-tune.
- **Domain generalization methods (DGMs):** MOSE (Qin et al., 2020) is a multi-domain mixture-of-experts (MoE) model, and each domain has its specific head. EDDFN (Silva et al., 2021) preserves domain-specific and domain-shared knowledge. MDFEND (Nan et al., 2021) utilizes a Domain Gate to select useful experts of MoE. CANMD (Yue et al., 2022) performs label shift correction and contrastive learning. MetaAdapt (Yue et al., 2023) adopts a meta-learning approach for domain-adaptive few-shot misinformation detection.
- **Retrieval method according to other types of embeddings:** We use the last_hidden_state of RoBERTa and another popular sentiment model (i.e. Sentibert (Yin et al., 2020)) as semantic and another kind of sentiment representation of each sentence respectively, then apply the same process of RAEmoLLM to deploy the ablation experiment.
- **Zero-shot and few-shot methods:** We also develop experiments of zero-shot method (LLMs-zs), randomly sample examples without using affective information (LLMs-random), and randomly sample examples with explicit Vreg information (LLMs-random-addexpl) for baselines.

3.2 Evaluation Metric

Misinformation detection is typically regarded as a classification task, therefore we employ a variety of metrics—Accuracy, Precision, Recall, and F1 for evaluation (Su et al., 2020) (All metrics use the weighted variant).

3.3 Results

We evaluate RAEmoLLM framework on one Nvidia Tesla A100 GPU with 80GB of memory. The max length of new tokens is 256 and do_sample is False. Others all use the default setting in the "model.generate"[6] package. We firstly select the instruction data based on Vreg to test the effectiveness of the RAEmoLLM framework on different LLMs. The result is the overall performance, which means that in AMTCele and PHEME, every domain is considered as the target domain test set, and the overall result is the performance of the combination of each domain test set. For Gemma series, Llama series and Vicuna series, we only show the best overall performing one in the table. In this section, we will be discussing results exclusively based on the F1 score. We firstly compare the RAEmoLLM framework with various baseline methods (e.g. PLMs, domain generalization methods, zero-shot, and few-shot methods) at Sec. 3.3.1. The ablation study of each module is conducted at Sec. 3.3.2. We subsequently compare the results on the data retrieved based on different affective information at Sec. 3.3.3.

[4] https://openai.com/
[5] https://llama.meta.com/llama3/
[6] https://huggingface.co/docs/transformers/en/main_classes/text_generation

	AMTCele				PHEME				COCO			
Model	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
BERT	0.5414	0.5453	0.5414	0.5322	0.7214	0.7203	0.7214	0.7208	0.7288	0.7510	0.7288	0.6356
RoBERTa	0.5678	0.7228	0.5678	0.4730	0.7199	0.7213	0.7199	0.7204	0.7328	0.7851	0.7328	0.6388
MDFEND	0.5878	0.5934	0.5878	0.5815	0.5796	0.6425	0.5796	0.5829	0.7988	0.7939	0.7988	0.7793
EDDFN	0.7041	0.7313	0.7041	0.6951	0.7004	0.6925	0.7004	0.6816	0.7116	0.5064	0.7116	0.5917
MOSE	0.5031	0.5051	0.5031	0.4482	0.7135	0.7130	0.7135	0.6890	0.7198	0.7335	0.7198	0.6162
CANMD	0.6296	0.6650	0.6296	0.6086	0.7382	0.7338	0.7382	0.7346	0.7291	0.7324	0.7291	0.6441
MetaAdapt	0.6429	0.6564	0.6429	0.6350	0.6193	0.6804	0.6193	0.6230	0.5186	0.7267	0.5186	0.5222
Mistral-7b-zs	0.7020	0.7346	0.7020	0.6926	0.5897	0.6491	0.5897	0.5936	0.3686	0.7050	0.3686	0.4673
Mistral-7b-random	0.7082	0.7768	0.7082	0.6889	0.6177	0.6334	0.6177	0.6227	0.7128	0.7455	0.7128	0.7287
Mistral-7b-random-addexpl	0.6337	0.7050	0.6337	0.5988	0.5804	0.6177	0.5804	0.5870	0.6802	0.7245	0.6802	0.7010
Mistral-7b-Vreg	0.7469	0.7748	0.7469	0.7404	0.6760	0.6837	0.6760	0.6788	0.7779	0.8031	0.7779	0.7898
Mistral-7b-Vreg-addexpl	**0.7735**	**0.7822**	**0.7735**	**0.7717**	**0.6921**	**0.6919**	**0.6921**	**0.6920**	**0.7814**	**0.8053**	**0.7814**	**0.7931**
Gemma-2b-zs	0.4153	0.4568	0.4153	0.3815	0.3606	0.5113	0.3606	0.2303	0.3302	0.4572	0.3302	0.3835
Gemma-2b-random	0.4980	0.4997	0.4980	0.4649	0.4269	0.5799	0.4269	0.3575	0.4477	0.6336	0.4477	0.4816
Gemma-2b-random-addexpl	0.4929	0.4928	0.4929	0.4927	**0.5914**	0.5777	**0.5914**	0.5820	0.6221	0.6164	0.6221	0.5587
Gemma-2b-Vreg	**0.6235**	**0.6298**	**0.6235**	**0.6213**	0.4361	**0.5953**	0.4361	0.3708	0.5302	**0.7326**	0.5302	0.5814
Gemma-2b-Vreg-addexpl	0.5847	0.6190	0.5847	0.5525	0.5875	0.5846	0.5875	**0.5859**	**0.6767**	0.6932	**0.6767**	**0.5990**
Llama3.2-1b-zs	0.4796	0.4841	0.4796	0.4801	0.5549	0.4480	0.5549	0.4712	0.5826	0.5997	0.5826	0.5385
Llama3.2-1b-random	0.5398	0.5483	0.5398	0.5222	0.3949	0.4831	0.3949	0.3417	0.7116	0.5064	0.7116	0.5917
Llama3.2-1b-random-addexpl	0.4867	0.4868	0.4867	0.4782	0.4118	0.4821	0.4118	0.3996	0.7116	0.5064	0.7116	0.5917
Llama3.2-1b-Vreg	0.6173	0.6360	0.6173	0.6065	0.6254	0.6432	0.6254	0.6307	0.7233	0.7640	0.7233	0.6242
Llama3.2-1b-Vreg-addexpl	**0.6429**	**0.6460**	**0.6429**	**0.6438**	**0.6473**	**0.6831**	**0.6473**	**0.6535**	**0.7372**	**0.7718**	**0.7372**	**0.6545**
ChatGPT-zs	**0.7265**	0.7420	**0.7265**	**0.7221**	0.5236	0.6551	0.5236	0.5032	0.7860	0.7920	0.7860	0.7551
ChatGPT-random	0.6990	0.7475	0.6990	0.6835	0.6173	0.6539	0.6173	0.6234	0.7616	0.7782	0.7616	0.7079
ChatGPT-random-addexpl	0.6959	0.7193	0.6959	0.6876	0.6092	0.6584	0.6092	0.6144	0.7651	0.7824	0.7651	0.7174
ChatGPT-Vreg	0.6745	0.7366	0.6745	0.6516	**0.6370**	0.6681	**0.6370**	**0.6429**	**0.8151**	**0.8249**	**0.8151**	**0.7925**
ChatGPT-Vreg-addexpl	0.7163	**0.7628**	0.7163	0.7032	0.6318	**0.6762**	0.6318	0.6372	0.8012	0.8068	0.8012	0.7772
GPT4o-zs	0.8816	0.8856	0.8816	0.8813	0.6170	0.6398	0.6170	0.6228	0.7837	0.8150	0.7837	0.7396
GPT4o-random	0.8776	0.8850	0.8776	0.8770	0.6739	0.6830	0.6739	0.6771	0.8291	0.8526	0.8291	0.8090
GPT4o-random-addexpl	0.8724	0.8824	0.8724	0.8716	0.6559	0.6693	0.6559	0.6601	0.8337	0.8527	0.8337	0.8158
GPT4o-Vreg	**0.8888**	**0.8934**	**0.8888**	**0.8884**	0.7004	0.6983	0.7004	0.6992	**0.8477**	**0.8627**	**0.8477**	**0.8326**
GPT4o-Vreg-addexpl	0.8847	0.8912	0.8847	0.8842	**0.7155**	**0.7170**	**0.7155**	**0.7162**	0.8419	0.8605	0.8419	0.8242
Vicuna-7b-zs	0.5490	0.5545	0.5490	0.5384	0.4378	0.6502	0.4378	0.3542	0.2942	0.7054	0.2942	0.1592
Vicuna-7b-random	0.5837	0.5872	0.5837	0.5806	0.4073	0.6116	0.4073	0.3017	0.7070	0.6037	0.7070	0.5928
Vicuna-7b-random-addexpl	0.5622	0.6040	0.5622	0.5206	0.5334	0.5849	0.5334	0.5423	0.7023	0.5063	0.7023	0.5884
Vicuna-7b-Vreg	0.6000	0.6069	0.6000	0.6023	0.4512	**0.6549**	0.4512	0.3821	**0.7837**	**0.7999**	**0.7837**	0.7471
Vicuna-7b-Vreg-addexpl	**0.6316**	**0.6680**	**0.6316**	**0.6248**	**0.6065**	0.6145	**0.6065**	**0.6105**	0.7756	0.7956	0.7756	**0.7501**

Table 3: Overall results on three datasets. "zs" denotes the zero-shot method. "random" denotes randomly sample four examples without using affective information. "random-addexpl" denotes adding explicit Vreg information for the random sample examples. "Vreg" denotes retrieving four examples based on Vreg information using Template 1. "Vreg-addexpl" denotes adding explicit Vreg information using Template 2.

3.3.1 Comparison with baselines

(1) Comparison with PLMs and other domain generalization methods: We can observe that most LLMs with RAEmoLLM framework outperform fine-tuned RoBERTa, BERT, and DGMs on AMTCele and COCO datasets, but they slightly underperform fine-tuned models and some DGMs in the PHEME dataset. One possible reason is that in cross-domain misinformation detection tasks, the fine-tuning method may perform better for simple short-text discrimination problems in the large-scale dataset (e.g. PHEME). However, they may not be suitable for long texts (e.g. AMTCele) or complex tasks (e.g. intent recognition in COCO), especially in small datasets. We can see that the current DGMs do not have stable performance on different datasets, although they have complex structures. And their results are lower than the best performance of LLMs with the RAEmoLLM framework in most cases.

(2) Comparison with zero-shot method (LLMs-zs), random few-shot methods (LLMs-random, LLMs-random-addexpl): From Table 3, we can observe that the RAEmoLLM framework largely increases the LLMs with zero-shot method in most cases and performs better than the random few-shot methods (For random few-shot, the largest increase in AMTCele is Gemma-2b (+15.64%), in PHEME is Llama3.2-1b (+31.18%), and in COCO is Vicuna-7b (+15.73%)). The results of LLMs-random-addexpl show that simply applying explicit information has little effect in most cases[7]. A special case is that in the AMTCele dataset, GPT-4o and ChatGPT perform well in zero-

[7]For Llama3.2-1b in COCO, both the random and random-addexpl variants predict all items as conspiracy category, resulting in the same results.

shot settings, with ChatGPT even surpassing other few-shot methods. One possible reason is that the AMTCele dataset is collected from fact-checking websites, and ChatGPT's and GPT-4o's training set includes the data and can effectively utilize this information. One example is shown in Table 10.

Table 7 and Table 8 in Appendix C present the performance of Mistral-7b on each domain on AMTCele and PHEME separately. It can be observed that Mistral-7b with RAEmoLLM framework overtakes Mistral with zero-shot and few-shot methods in most domains except for the prince domain in PHEME, which has significantly imbalanced data. Additionally, we also conduct some special cases analysis in Appendix D.

3.3.2 Ablation analysis of each module

(1) Index Construction Module (retrieval based on different information): From Table 3, we can observe retrieval based on affective information (LLMs-Vreg, LLMs-Vreg-addexpl) overtake non-retrieval methods (i.e. random few-shot methods (LLMs-random, LLMs-random-addexpl)). From Table 4, we can observe that the RAEmoLLM framework achieves the best results compared to other types of embeddings, which indicates the effectiveness of Vreg embedding.

	AMT	PHEME	COCO
Mistral-7b-Vreg	0.7404	0.6788	0.7898
Mistral-7b-Vreg-addexpl	0.7717	0.6920	0.7931
Mistral-7b-semantic	0.6904	0.6718	0.7771
Mistral-7b-sentibert	0.6984	0.6663	0.7687

Table 4: F1 score of retrieval using different kinds of embeddings. "semantic" denotes retrieval based on RoBERTa.

(2) Retrieval Module (different numbers of retrieval examples): Table 5 presents the F1 score of retrieval of different numbers of examples based on Vreg (we only tested 16 examples in the AMTCele dataset due to its long text). From the table, it can be observed that increasing the retrieval examples does not consistently improve the model's performance, and it may even lead to a decline in its performance (e.g. Vreg-addexpl in COCO). One possible reason is that when the model has multiple examples as references, it needs to consider a large amount of information comprehensively, which depends on the model's capability. Another reason we can infer from Table 14. For the three datasets, the p-values in retrieval top 4 examples are all zero. However, as the number of retrieval examples increases, the second p-values in AMTCele and the first p-value in COCO dataset also gradually increase. This indicates that the retrieved content may come from another category or unrelated examples, thereby affecting the model's judgment ability. Therefore, when employing retrieval augmentation techniques, it is not just about blindly increasing the number of examples, but rather selectively choosing the most useful examples.

Datasets	methods	4	8	16	32	64
AMTCele	Random	0.6889	0.7006	0.6287	-	-
	Vreg	0.7404	0.7395	0.7271	-	-
	Vreg-addexpl	0.7717	0.7611	0.7710	-	-
PHEME	Random	0.6227	0.6253	0.6268	0.6400	0.6353
	Vreg	0.6788	0.6856	0.6830	0.6910	0.7031
	Vreg-addexpl	0.6920	0.6949	0.6979	0.6979	0.6990
COCO	Random	0.7287	0.7534	0.7442	0.7628	0.7541
	Vreg	0.7898	0.7842	0.7854	0.8172	0.7993
	Vreg-addexpl	0.7931	0.7208	0.7499	0.7600	0.7475

Table 5: F1 score of Mistral-7b with retrieval of different numbers of examples based on Vreg.

(3) Inference Module (different templates and different base LLMs): We can see LLMs with explicit affective information based on Template 2 (i.e. LLMs-Vreg-expl) exceed LLM without explicit affective information based on Template 1 (i.e. LLMs-Vreg) in most cases. For LLMs-zs and LLMs-random, different base models show significant performance differences. GPT-4o performs the best, followed by ChatGPT and Mistral-7b, while the Gemma-2b model has the lowest score. After using RAEmoLLM framework, the difference between different modules becomes narrowing (e.g. Mistral-7b has achieved or even surpassed the performance of ChatGPT.)

Based on the analysis above, we can conclude that retrieval based on implicit affective information and adding explicit affective information through Template 2 is the most effective way to enhance the LLMs' performance in using Vreg affective cases. The number of retrieval examples seems to have little impact. The LLMs focus on the most relevant examples.

Table 3 shows that Mistral-7b has the best performance among open-sourced LLMs. We choose Mistral-7b to conduct the following experiments.

3.3.3 Results on the data retrieved based on different affective information

Figure 2 presents the results of retrieval with different affective embeddings. For retrieval using affective regression information (i.e. Vreg, EIreg), it is evident that adding explicit affective information (*affect-addexpl*) method can improve the per-

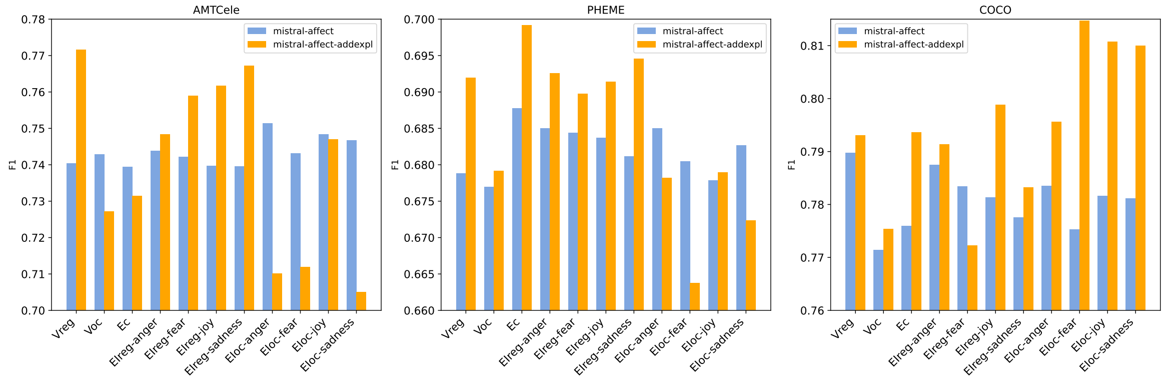

Figure 2: Results of Mistral-7b based on different affective information on three datasets. “affect” denotes retrieving four examples based on one affective information using Template 1. “affect-addexpl” denotes adding explicit affective information using Template 2.

formance compared to solely relying on implicit affective information retrieval (*affect*). However, when using affective classification information (e.g. EIoc in AMTCele and PHEME), adding explicit affective information may confuse the model. In the COCO dataset, all the *affect-addexpl* method outperforms *affect* except for EIreg-fear. Regarding the *affect-addexpl* method, in AMTcele, we can see the results retrieval based on Vreg are best, followed by EIreg-sadness and EIreg-joy. And the final three rankings are retrieved based on EIoc-anger, fear, and sadness. It seems that affective intensity and strength are more suitable for cross-domain fake news detection tasks. In PHEME, retrieval based on Ec exhibits the highest performance, with the Vreg and EIreg series closely trailing behind. While the last few are the EIoc series, which may suggest that a coarse-grained emotional intensity classification is not suitable for rumour detection. However, it is the opposite in the conspiracy theory dataset. In COCO, the performance of retrieval based on the EIoc series is better than that based on the EIreg series. This could be attributed to the dataset’s focus on COVID-19 topic, which may elicit more consistent emotional expressions among individuals.

4 Conclusion and Future Work

In this paper, we propose RAEmoLLM, the first RAG framework to address cross-domain misinformation detection using in-context learning based on affective information. We introduce the three modules of RAEmoLLM. We also conduct a comprehensive affective analysis for three public misinformation datasets. We evaluate the performance of RAEmoLLM on the three misinformation benchmarks based on various LLMs. The results show that RAEmoLLM can significantly improve LLMs compared to the zero-shot method and other few-shot methods, which illustrates the effectiveness of RAEmoLLM. We also conduct an ablation analysis of each module and analyze the performance of retrieval based on different affective information, which provide a foundation for further improvements in the future.

In the future, we will explore the application of multimodal affective information in the task of detecting misinformation. We will also evaluate the application of the RAEmoLLM framework in other fields (e.g. mental health and finance). In addition to affective information, there are many other influencing factors in misinformation, such as stance and topic. We will combine sentiments and emotions with other features to construct a more robust retrieval database. Furthermore, the retrieval process can be slowed down by a large amount of data. In the future, we will also explore more efficient retrieval methods.

5 Limitations

Due to restricted computational resources, we only carried out inference of 1B, 2B, 7B, 8B, 13B, and 33B open-sourced LLMs. As such, we have not considered how the use of larger or different model architectures may potentially impact upon performance in cross-domain misinformation detection tasks.

Though achieving outstanding performance, RAEmoLLM still bears limitations. Firstly, for domain data with imbalanced distribution, RAEmoLLM performs worse compared to zero-shot methods (e.g. prince domain in PHEME). The special cases analysis in Appendix D also illustrates that in the imbalanced datasets, the retrieval in RAEmoLLM will be influenced for some special cases. Therefore, further exploration is needed to address such issues. Secondly, in the PHEME dataset, RAEmoLLM performs worse than fine-tuning methods without emotional information. This indicates that for simple tasks with shorter texts, the model still struggles to effectively balance textual features and emotional information.

Acknowledgments

This work is supported by the Turing Scheme Fund, the Manchester-Melbourne-Toronto Research Fund, the Centre for Digital Trust and Society at the University of Manchester, and the New Energy and Industrial Technology Development Organization. This work is also supported by the scholar award from the Department of Computer Science at the University of Manchester.

References

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853. IEEE.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Arjun Choudhry, Inder Khatri, Arkajyoti Chakraborty, Dinesh Vishwakarma, and Mukesh Prasad. 2022. Emotion-guided cross-domain fake news detection using adversarial domain adaptation. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 75–79.

Carmela Comito, Francesco Sergio Pisani, Erica Coppolillo, Angelica Liguori, Massimo Guarascio, and Giuseppe Manco. 2023. Towards self-supervised cross-domain fake news detection.

Nicholas Della Giustina. 2023. Misinformation and its effects on individuals and society from 2015-2023: A mixed methods review study.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diwen Dong, Fuqiang Lin, Guowei Li, and Bo Liu. 2022a. Sentiment-aware fake news detection on social media with hypergraph attention networks. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2174–2180. IEEE.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022b. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. 2017. Emotion analysis: A survey. In *2017 international conference on computer, communications and electronics (COMPTELIX)*, pages 397–402. IEEE.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Suzanne Keen. 2006. A theory of narrative empathy. *Narrative*, 14(3):207–236.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, pages 1–42.

Guanghua Li, Wensheng Lu, Wei Zhang, Defu Lian, Kezhong Lu, Rui Mao, Kai Shu, and Hao Liao. 2024. Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors. *arXiv preprint arXiv:2403.09747*.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. *arXiv preprint arXiv:2309.13205*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. 2024a. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024b. Conspemollm: Conspiracy theory detection using an emotion-based large language model. In *ECAI 2024*, pages 4649–4656. IOS Press.

Zhiwei Liu, Kailai Yang, Tianlin Zhang, Qianqian Xie, Zeping Yu, and Sophia Ananiadou. 2024c. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. *arXiv preprint arXiv:2401.08508*.

Zhiwei Liu, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu, and Sophia Ananiadou. 2024d. Emotion detection for misinformation: A review. *Information Fusion*, page 102300.

Quanyu Long, Wenya Wang, and Sinno Jialin Pan. 2023. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. *arXiv preprint arXiv:2311.11551*.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401.

Zhen Qin, Yicheng Cheng, Zhe Zhao, Zhe Chen, Donald Metzler, and Jingzheng Qin. 2020. Multitask mixture of sequential experts for user activity streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3083–3091.

Tanik Saikh, Arkadipta De, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A deep learning approach for automatic detection of fake news. *arXiv preprint arXiv:2005.04938*.

Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669.

Jiao Shi, Xin Zhao, Nan Zhang, Yu Lei, and Lingtong Min. 2023. Rough-fuzzy graph learning domain adaptation for fake news detection. *IEEE Transactions on Computational Social Systems*.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1-2):1–13.

Wei Tang, Zuyao Ma, Haifeng Sun, and Jingyu Wang. 2023. Learning sparse alignments via optimal transport for cross-domain fake news detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *ACM Multimedia 2024*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2024. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36.

Yi Wu, Ziqiang Li, Chaoyue Wang, Heliang Zheng, Shanshan Zhao, Bin Li, and Dacheng Tao. 2024. Domain re-modulation for few-shot generative domain adaptation. *Advances in Neural Information Processing Systems*, 36.

Shangqing Xu and Chao Zhang. 2024. Misconfidence-based demonstration selection for llm in-context learning. *arXiv preprint arXiv:2401.06301*.

Xin Xu, Yue Liu, Panupong Pasupat, Mehran Kazemi, et al. 2024. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2423–2433.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5628–5643.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5223–5239.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023a. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xuewen Zhang, Yaxiong Pan, Xiao Gu, and Gang Liang. 2023b. Sentiment analysis-based social network rumor detection model with bi-directional graph convolutional networks. In *International Conference on Computer Application and Information Security (ICCAIS 2022)*, volume 12609, pages 463–469. SPIE.

A Related Work

A.1 Misinformation detection

Cross-domain misinformation detection: Cross-domain misinformation detection refers to identifying and detecting misleading or false information across different domains or sources. Comito et al. (2023) propose a deep learning-based architecture to mitigate this problem by yielding high-level cross-domain features. Tang et al. (2023) design one framework to learn transferable features across domains by aligning the source and target news using Optimal Transport techniques. Shi et al. (2023) develop a rough-fuzzy graph learning framework that uses representations of cross-domain sample uncertainty structural information, and captures shared general features across domains. Tong et al. (2024) integrate domain embeddings and attention mechanisms for domain-specific knowledge extraction and combine techniques to obtain multi-domain and multi-modal information. Nan et al. (2021) adopt domain gates to aggregate multiple representations extracted by a mixture of experts (MoE) for fake news detection. Silva et al. (2021) jointly leverage domain-specific and cross-domain knowledge and introduces an unsupervised technique to train a multi-domain fake news detection model. Yue et al. (2022) propose a contrastive adaptation network, which leverages pseudo-labeling to generate target examples and design a label correction component to solve label shift problems. Yue et al. (2023) develop a domain-adaptive few-shot method based on meta-learning, which adopts limited target examples to provide feedback and guide knowledge transfer from the source domain to the target domain. However, these methods require complex structures and fine-tuning strategies.

Retrieval augmented misinformation detection: Retrieval augmented generation (RAG) combines LLMs with retrieval systems to utilize external knowledge, enabling models to generate more accurate content. Xuan et al. (2024) leverage LVLM intuition and reasoning capabilities to enhance the accuracy of multimodal misinformation by retrieving external knowledge. Yue et al. (2024) collect supporting evidence from scientific sources and generate responses for combating misinformation online based on this evidence. Cheung and Lam (2023) adopt external, most up-to-date information available on the Internet to bridge the knowledge gap in an LLM to enhance fake news detection performance. Li et al. (2024) employ a multi-round retrieval strategy, which can extract key evidence from web sources for claim verification. These findings demonstrate the effectiveness of RAG technology in detecting misinformation.

Affect-based misinformation detection: Emotion and sentiment are important features for misinformation detection (Liu et al., 2024d). Zhang et al. (2023b) combine the use of semantic and sentiment information, along with propagation information for rumour detection. Dong et al. (2022a) design a sentiment-aware hyper-graph attention network for fake news detection. Liu et al. (2024b) develop a conspiracy theory detection LLM by fine-tuning EmoLLaMA (Liu et al., 2024c). Choudhry et al. (2022) utilize emotional information for fake news detection based on an adversarial learning structure. Unfortunately, these works either have complex structural designs or fine-tuned models, which require significant time and computational resources. RAEmoLLM in this paper applies the ICL method based on retrieving demonstration examples through affective information, which has a simple structure and does not involve fine-tuning.

A.2 In-context learning

In-context learning (ICL) is a specific prompting engineering method, in which the task demonstrations are included in prompts for LLMs learning (Xu et al., 2024). Wang et al. (2023) develop a framework to provide high-quality context examples for LLMs, which firstly evaluate the quality of candidate examples through a reward model, and then conduct knowledge distillation to train a dense retriever. Wang et al. (2024) introduce an algorithm that utilizes a small LM to select the best demonstrations from a set of annotated data, and subsequently expand these demonstrations to larger LMs. Liu et al. (2024a) develop in-context curriculum learning, a simple but helpful demonstration ordering method for ICL that gradually increases the complexity of prompt demonstrations. Xu and Zhang (2024) propose in-context reflection to strategically select demonstrations that reduce the discrepancy between the LLM's outputs and the actual input-output mappings. Long et al. (2023) propose a retrieval-enhanced language model to address cross-domain problems, in which they train language models by learning both target domain distribution and the discriminative task signal simultaneously with the augmented cross-domain in-context examples. Inspired by these works, we propose the RAEmoLLM.

B Task Prompt and Instruction Example

For AMTCele, we utilize *"Determine whether the target text is 0. Fake or 1. Legit."* For PHEME, we employ *"Determine if the target text is 0. non-rumours or 1. rumours."* For COCO, we apply *"Classify the text regarding COVID-19 conspiracy theories or misinformation into one of the following three classes: 0. Unrelated. 1. Related (but not supporting). 2. Conspiracy (related and supporting)."* Here we keep the *0. Unrelated* category to test the robustness of the LLM by increasing the complexity of the task.

Table 6 presents a specific instruction example.

Task: Determine if the target text is 0. non-rumours or 1. rumours.
Target text: UPDATE: Reports of 1 more shooter being SHOT. This is in addition to one shot and killed earlier in Parliament Hill #OttawaShooting. Sentiment intensity: 0.234.
Here are a few examples retrieved through sentiment intensity:
Text: UPDATE: Reports gunman says four devices are located around Sydney. Security response underway. Police calling for calm. #9News. Sentiment intensity: 0.429. The label of this text: 1. rumours.
Text: JUST IN: Police confirm to ABC there is a second hostage situation underway in eastern Paris. Sentiment intensity: 0.328. The label of this text: 1. rumours.
Text: UPDATE: There are reports police have discovered the identity of the lone gunman, with the #SydneySiege in its sixth hour. #9News Sentiment intensity: 0.435. The label of this text: 1. rumours.
Text: JUST IN: A separate shooting and hostage situation at a supermarket in eastern Paris has been reported ... developing. Sentiment intensity: 0.236. The label of this text: 1. rumours.
According to the above information, the label of target text:

Table 6: An example in the PHEME instruction dataset.

C The results from different domains in the AMTCele and PHEME datasets. (Table 7 and 8)

D Special cases analysis

Misinformation and true information often convey different affective information (as shown in Table 2 and Table 13). For example, fake news and conspiracy theories tend to evoke more negative sentiments and emotions (e.g. anger or fear) and less joy. However, these results are based on statistics derived from the entire dataset. The special cases need to be analyzed. We investigate some special cases retrieved based on EIoc. The results are listed in Table 9.

For AMTCele, we investigate cases where fake news lacks anger or exhibits higher levels of joy, as well as cases where legit news displays higher levels of anger or lacks joy. We can see that the examples retrieved are mostly of the same category as the target, and their results have not been greatly influenced. For PHEME and COCO, we calculate statistics on cases of rumour and conspiracy without fear or exhibiting higher levels of joy (we do not report conspiracy with higher joy due to its low occurrence), as well as cases where non-rumour and related display higher levels of fear or without joy. We can see that the results for rumours in PHEME and related in COCO are poor. The most likely reason is due to the imbalance of categories in the original data, and these special cases are in the minority. This has resulted in the retrieval of more data from the larger category in original datasets, causing the model to learn less useful information and ultimately affecting the final results.

E Data leakage example in AMTCele (Table 10)

F Comparison of time consumption between RAEmoLLM and fine-tuning methods (Table 11)

We take the PHEME dataset (6425 items) as an example to compare the time consumption between RAEmoLLM (applying ChatGPT as the base

	biz		edu		entmt		polit		sports		tech		celebrity	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT	0.5975	0.5930	0.5725	0.5436	0.5800	0.5610	0.5450	0.5180	0.5525	0.5293	0.5650	0.5409	0.5152	0.5039
Mistral-7b-zs	0.7250	0.7135	0.8000	0.7954	0.7625	0.7595	0.5750	0.5157	0.7750	0.7714	0.6000	0.5442	0.6980	0.6925
Mistral-7b-random	0.7375	0.7218	0.6625	0.6191	0.7375	0.7251	0.5500	0.4357	0.6875	0.6761	0.5625	0.4589	0.7580	0.7489
Mistral-7b-Vreg	0.7750	0.7656	0.8250	0.8222	0.8250	0.8222	0.6125	0.5706	0.8125	0.8089	0.7250	0.7067	0.7320	0.7275
Mistral-7b-Vreg-addexpl	0.8000	0.7968	0.8625	0.8620	0.8500	0.8496	0.6625	0.6423	0.8375	0.8373	0.8625	0.8607	0.7360	0.7346

Table 7: The results from different domains in the AMTCele dataset

	sydneysiege		ottawashooting		charliehebdo		ferguson		germanwings		prince		putinmissing		gurlitt		ebola	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT	0.7463	0.7418	0.7497	0.7490	0.7971	0.8113	0.7053	0.7147	0.7275	0.7260	0.1296	0.1985	0.5866	0.5297	0.5391	0.4949	0.5714	0.7220
Mistral-7b-zs	0.6536	0.6552	0.6506	0.6504	0.6075	0.6407	0.4051	0.4146	0.6716	0.6638	0.7382	0.8344	0.5546	0.4807	0.4420	0.4389	0.4286	0.6000
Mistral-7b-random	0.6822	0.6838	0.5719	0.5232	0.6946	0.7153	0.4506	0.4653	0.6652	0.6646	0.6395	0.7636	0.5378	0.4569	0.5362	0.4225	0.3571	0.5263
Mistral-7b-Vreg	0.7215	0.7195	0.6652	0.6596	0.7335	0.7521	0.5818	0.6102	0.7143	0.7139	0.5451	0.6881	0.6008	0.5716	0.4928	0.4514	0.5000	0.6667
Mistral-7b-Vreg-addexpl	0.7437	0.7403	0.6753	0.6683	0.7431	0.7613	0.6527	0.6655	0.7036	0.7033	0.4592	0.6128	0.6050	0.6023	0.4348	0.4308	0.4286	0.6000

Table 8: The results from different domains in the PHEME dataset

Datasets	EIoc	num	F1	mean num of retrieval	
				legit/non-rum/related	fake/rumour/consp
AMT	fake anger=0	218	0.8152	1.0780	2.9220
	legit anger=2/3	29	0.9643	2.2414	1.7586
	fake joy=2/3	14	0.6667	1.5000	2.5000
	legit joy=0	304	0.8571	2.1217	1.8783
PHEME	non-rum fear=2/3	446	0.6978	2.4776	1.5224
	rumour fear=0	1039	0.3804	2.4658	1.5342
	non-rum joy=0	3795	0.8949	2.9057	1.0943
	rumour joy=2/3	25	0.2759	3.7600	0.2400
COCO	related fear=2/3	47	0.5538	2.0426	1.9574
	consp fear=0	171	0.9073	0.9708	3.0292
	realted joy ==0	246	0.7607	2.2927	1.7073

Table 9: Special cases retrieval based on EIoc. "num" denotes number. "non-rum" denotes non-rumours. "consp" denotes conspiracy. The "0", "2", and "3" in the EIoc column represent "no", "moderate", and "high" emotional intensity.

Prompt	News content: Alex Jones Apologizes for Promoting 'Pizzagate' Hoax Alex Jones a prominent conspiracy theorist and the host of a popular right-wing radio show . . . **Where does this news come from?**
GPT-3.5-trubo	This news comes from a New York Times article
GPT-4o	The news about Alex Jones apologizing for promoting the Pizzagate conspiracy theory likely comes from a reputable news source such as The New York Times, The Washington Post, or another major media outlet that covers significant events and developments involving public figures and misinformation. The details provided in your question, such as the mention of "The Alex Jones Show" and the Pizzagate conspiracy, align with coverage typically found in mainstream news articles addressing misinformation and its impact.

Table 10: Data leakage example in AMTCele

model) and the fine-tuning method (BERT). From Table 11, it can be observed that RAEmoLLM will consume about 122s to construct the retrieval database (Obtain embeddings: 72s, Retrieval examples: 50s) and 208s to obtain the affective labels. For fine-tuning methods, we take BERT as an example. The current time consumed (Training each epoch: 3906s) by BERT was measured based on a single set of hyperparameters (e.g., batch size and learning rate). In practice, fine-tuning methods may require more time and effort to optimize hyperparameters. Overall, the RAEmoLLM process is simpler and more efficient.

G Affective analysis

G.1 Five types of affective information

(1) *Emotion intensity (EIreg):* For each of four different emotions (anger, fear, joy and sadness), assign a score between 0 and 1 to represent the intensity of emotion of the text;

(2) *Emotion intensity classification (EIoc):* The text can be classified into one of four classes of the intensity of emotion (anger, fear, joy, sadness), i.e. no/low/moderate/high emotional intensity;

(3) *Sentiment (valence) strength (Vreg):* Assign a real-valued score between 0 (most negative) and 1 (most positive) to represent the sentiment intensity of the text.

(4) *Sentiment (valence) classification (Voc):* The text can be categorized into one of seven ordinal classes (i.e. *{very, moderately, slightly} negative, neutral, {slightly, moderately, very} positive*);

(5) *Emotion detection (Ec):* The text can be classified as 'neutral or no emotion' or as one, or more, of eleven given emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust).

	Obtain embeddings	Obtain labels	Retrieval Examples	Inference (time/item)
RAEmoLLM	71.68s	208s	50s	0.48s
	Train (time/epoch)	Inference (time/item)		
Bert	3906.31s	0.093s		

Table 11: Time consumption of RAEmoLLM (take ChatGPT as base model) and fine-tuning methods (task BERT as the example) based on the PHEME dataset.

G.2 Further Affective Analysis

We show the statistics values and distribution of labels and embeddings in this Section. In Figures 3 to Figure 8, the y-axis represents the distribution of labels within the intention class indicated on the x-axis. The affective analysis on COCO has been done by ConspEmoLLM (Liu et al., 2024b). The figures show that most fake information convey more negative sentiments/emotions and less positive emotions compared to real categories[8]. Figure 9 and Figure 10 present the 3D visualization of affective embeddings on AMTCele and PHEME respectively. Table 13 shows the statistics values of EIreg and Vreg on PHEME and COCO.

To explore the relationship between affective classification information and misinformation, we conduct a chi-squared significance test and create two categorical variables. One is the misinformation label (real and fake), and the other variable is affective information. For EIoc, we count the values for 0 (absence) and others (presence) of a certain emotion. For Voc, we count the values of 7 classes. For Ec, we count the number of instances that contain each of the 11 emotions individually. Assuming the null hypothesis that affective signals are independent of text truthfulness, the chi-squared test results in Table 12 show p-values close to 0, allowing us to reject the null hypothesis. Overall, affective classification signals are also statistically linked to the veracity of the news.

Table 14 shows statistics of different affective embeddings (i.e. last hidden layer of EmoLLaMA-chat-7B). We perform t-tests on the top-K cosine similarity within categories and the cosine similarity between categories. For example, "fake-legit" denotes computing the cosine similarity between each data point in the "fake" category and each data point in the "legit" category. We then selected the top-K similarity values and performed t-test on them. The t-value and p-value of the top-4 similarity values between "fake-legit" and "fake-fake" are -22.516 and 0, which demonstrates that the top 4 similar data retrieved based on cosine similarity within the "fake" category are highly likely to belong to the same "fake" category. We can see from the results in Table 14 that all affective information leads to the same conclusion in the top-4 scenarios[9]. We also visualize the data distribution reduced to 3 dimensions using PCA in Figures 9 and 10 in Appendix. It can be observed that different categories are clearly separated in the latent space. All the above demonstrate the close relationship between affective information and misinformation.

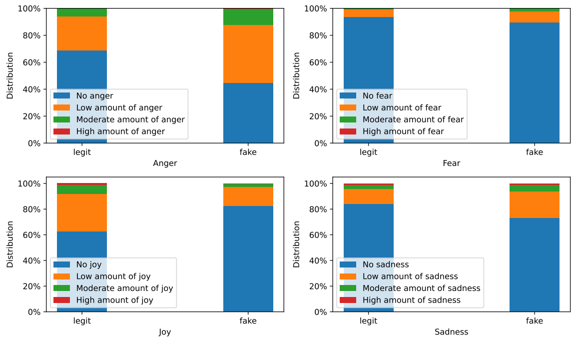

Figure 3: Emotion intensity classification on AMTCele

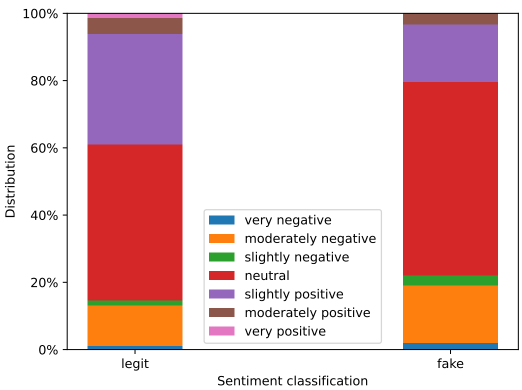

Figure 4: Sentiment classification on AMTCele

[8]Rumours are a complex category, encompassing true rumours, false rumours, and unverified rumours. Due to space constraints, a detailed analysis is not provided here. Nevertheless, different types convey different affective information.

[9]It should be noted that in Vreg, as the value of K increases, the second p-value in AMTCele and the first p-value in COCO dataset also gradually increase, which may affect the results. Therefore, we choose K to be 4. The analysis of different values of K can be found in Section 3.3.2.

	AMT			PHEME			COCO		
	EIoc	Voc	Ec	EIoc	Voc	Ec	EIoc	Voc	Ec
chi-squared statistic	131.16	46.07	69.40	197.98	146.14	499.48	76.31	25.09	61.50
p-value	3.60E-25	2.86E-08	5.78E-11	3.08E-39	5.07E-29	5.69E-101	7.76E-14	3.28E-04	1.88E-09

Table 12: Chi-squared statistics values of EIoc, Voc, Ec on different datasets.

Datasets	Affective	sub-emotion	non-rumours/related		rumours/conspiracy		t-test	
			mean	var	mean	var	t	p
PHEME	EIreg	Anger	0.4547	0.0102	0.4233	0.0075	12.7093	1.44E-36
		Fear	0.5337	0.0170	0.5632	0.0198	-8.5027	2.28E-17
		Joy	0.2134	0.0121	0.1817	0.0133	11.0177	5.58E-28
		Sadness	0.5215	0.0152	0.5177	0.0182	1.1442	0.2526
	Vreg		0.4331	0.0143	0.3842	0.0139	15.9786	2.18E-56
COCO	EIreg	Anger	0.5475	0.0088	0.5641	0.0068	-4.5211	6.43E-06
		Fear	0.5623	0.0097	0.6034	0.0077	-10.5568	1.56E-25
		Joy	0.1800	0.0111	0.1514	0.0075	7.2230	6.66E-13
		Sadness	0.4701	0.0098	0.4773	0.0073	-1.8808	0.0601
	Vreg		0.3961	0.0095	0.3973	0.0066	-0.3325	0.7395

Table 13: T-test statistics values of EIreg and Vreg on different datasets. The t-test is conducted between *non-rumours/related* and *rumours/conspiracy*.

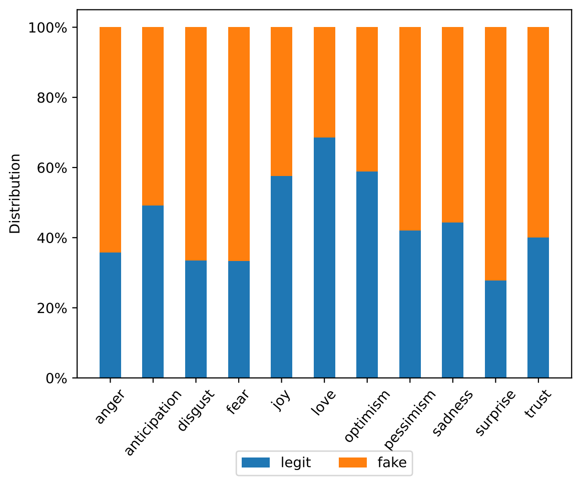

Figure 5: Emotion classification on AMTCele

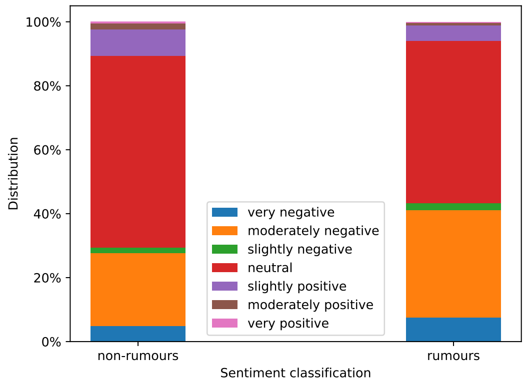

Figure 7: Sentiment classification on PHEME

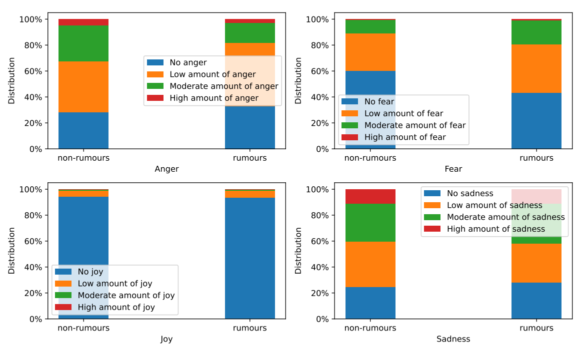

Figure 6: Emotion intensity classification on PHEME

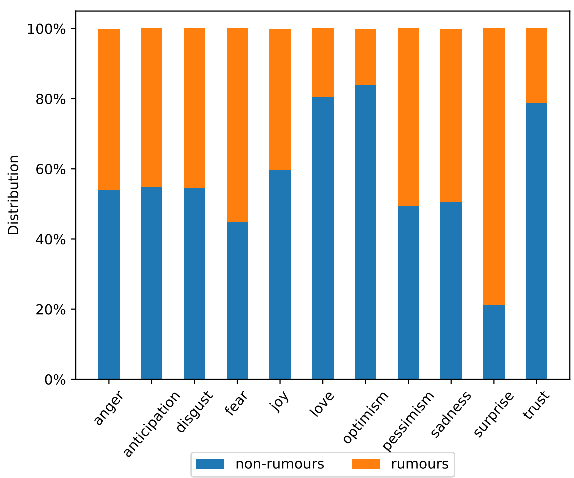

Figure 8: Emotion classification on PHEME

Datasets	Values	Vreg					Voc	Ec	EIreg				EIoc			
		top 4	top 8	top 16	top 32	top 64			anger	fear	joy	sadness	anger	fear	joy	sadness
AMTCele	fake-legit	0.791	0.771	0.753	0.736	0.718	0.852	0.812	0.801	0.801	0.801	0.801	0.840	0.840	0.840	0.840
	fake-fake	0.848	0.810	0.783	0.761	0.741	0.894	0.862	0.855	0.855	0.855	0.855	0.885	0.885	0.885	0.885
	t	-22.516	-14.875	-10.951	-8.976	-8.037	-20.550	-22.617	-22.434	-22.433	-22.462	-22.461	-22.260	-22.246	-22.267	-22.244
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	legit-fake	0.787	0.765	0.747	0.729	0.711	0.848	0.807	0.797	0.797	0.797	0.797	0.836	0.836	0.836	0.836
	legit-legit	0.841	0.798	0.768	0.743	0.721	0.886	0.856	0.848	0.848	0.848	0.848	0.877	0.877	0.877	0.877
	t	-21.568	-12.845	-8.052	-5.263	-3.452	-17.138	-21.024	-21.399	-21.387	-21.407	-21.396	-19.364	-19.328	-19.335	-19.315
	p	0.000	0.000	0.001	0.008	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
PHEME	nonr-rum	0.930	0.927	0.924	0.921	0.917	0.982	0.952	0.940	0.940	0.940	0.939	0.972	0.972	0.972	0.972
	nonr-nonr	0.957	0.946	0.938	0.932	0.927	0.989	0.971	0.963	0.963	0.963	0.963	0.983	0.983	0.983	0.983
	t	-75.127	-49.017	-35.035	-27.844	-24.327	-69.237	-78.344	-77.082	-77.231	-76.869	-78.103	-71.392	-71.732	-71.005	-72.538
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	rum-nonr	0.935	0.932	0.929	0.925	0.921	0.984	0.957	0.945	0.944	0.945	0.944	0.974	0.974	0.974	0.974
	rum-rum	0.961	0.950	0.942	0.935	0.928	0.990	0.974	0.966	0.966	0.967	0.966	0.984	0.984	0.984	0.984
	t	-58.813	-38.823	-27.206	-19.693	-14.156	-54.654	-58.600	-59.494	-59.637	-59.377	-60.266	-55.874	-56.306	-56.033	-56.759
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
COCO	rela-consp	0.873	0.870	0.866	0.861	0.856	0.955	0.905	0.885	0.885	0.886	0.885	0.936	0.936	0.937	0.936
	rela-rela	0.907	0.887	0.875	0.865	0.857	0.967	0.931	0.916	0.916	0.916	0.916	0.953	0.953	0.954	0.954
	t	-44.603	-23.007	-11.581	-5.437	-2.012	-37.288	-43.522	-44.744	-44.772	-44.253	-44.800	-38.201	-38.337	-37.684	-38.281
	p	0.000	0.093	0.428	0.457	0.312	0.004	0.000	0.000	0.000	0.001	0.000	0.001	0.001	0.002	0.002
	consp-rela	0.863	0.858	0.852	0.846	0.838	0.950	0.897	0.876	0.876	0.877	0.876	0.929	0.929	0.930	0.929
	consp-consp	0.911	0.891	0.878	0.868	0.859	0.968	0.933	0.919	0.919	0.920	0.920	0.954	0.954	0.955	0.954
	t	-74.176	-47.239	-33.132	-25.606	-21.079	-54.114	-69.563	-73.828	-73.876	-73.190	-73.709	-60.255	-60.393	-59.577	-60.204
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 14: Statistics values of cosine similarity between embeddings of different affective information on three datasets. Top K denotes retrieval top K examples. In addition to Vreg, the results of other affective information are all based on top 4. “A-B” represents the calculation of cosine similarity between each data point in A and each data point in B. Each element (i, j) in the resulting calculation represents the cosine similarity between the i-th vector in the A group embeddings and the j-th vector in the B group embeddings. The top 4 refers to selecting the four highest values from each row. The t-value and p-value represent the t-test results for the “A-B” results of the two lines above.

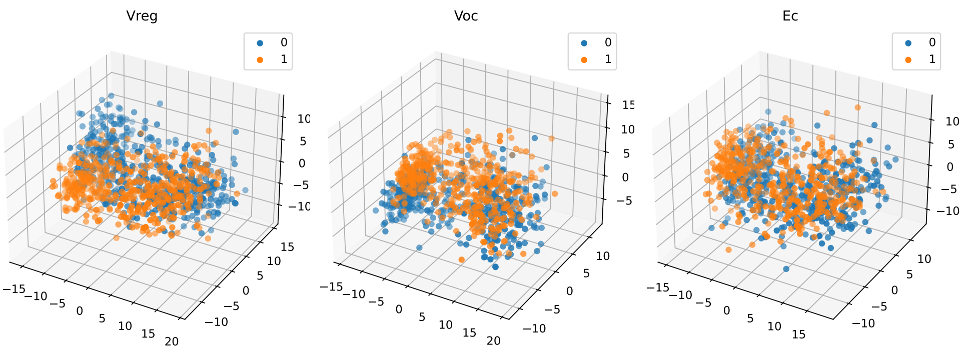

Figure 9: 3D visualization of affective embeddings on AMTCele. 0: Fake. 1: Legit

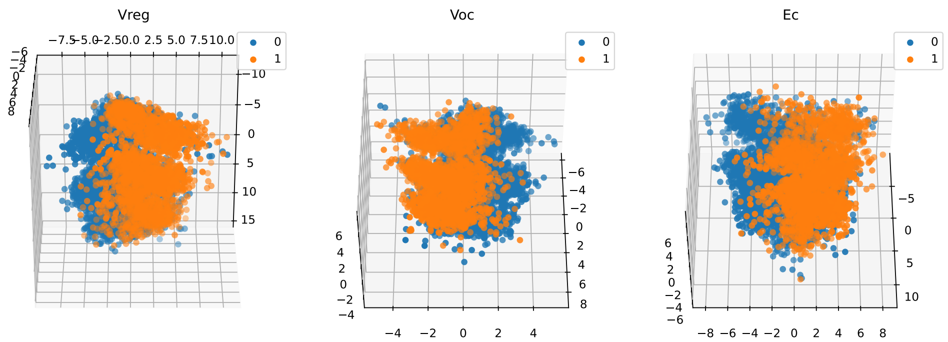

Figure 10: 3D visualization of affective embeddings on PHEME. 0: Non-rumours. 1: Rumours