# NoviceTrio at #SMM4H-HeaRD 2025: Leveraging LLMs and Ensemble Methods for Classifying Dementia Family Caregivers on Twitter

**Abir Naskar**[1], **Liuliu Chen**[1], **Jemima Kang**[1, 2], **Mike Conway**[1]

[1]School of Computing and Information Systems, University of Melbourne, Australia
[2]Melbourne School of Psychological Sciences, University of Melbourne, Australia
{abir.naskar, liuliu.chen, jemimak}@student.unimelb.edu.au, mike.conway@unimelb.edu.au

**Abstract**

Traditional interventions for dementia caregivers remain inaccessible and research has yet to explore the use of social media platforms to provide internet-based support. Task 3 of the 10th Social Media Mining for Health Workshop (SMM4H 2025) focuses on the task of binary classification of dementia family caregivers on Twitter. This paper presents a system that evaluates multiple pretrained language models, including BERT variants and LLMs, using ensembling and prompting strategies. Our LLM ensemble achieved the best performance with an F1 score of 0.957. We discuss the strengths and limitations of LLMs in this context.

## 1 Introduction

Health informatics has advanced rapidly in recent years (Ravì et al. 2016). Natural Language Processing (NLP) techniques have demonstrated strong abilities in understanding health-related problems and developing more efficient, accurate classification systems (Qiu et al. 2023). Furthermore, integrating social media data with NLP has been widely studied due to the rich, real-world insights it provides and its potential to enable timely identification and intervention (Moorhead et al. 2013).

Among all health problems, dementia has become a major concern, with over 55 million people affected worldwide in 2020, and is expected to almost double every 20 years (Alzheimer's Disease International 2015). Recent research has shown that Twitter is widely used by family caregivers of people with dementia to share experiences and seek support (Cheng, Liu, and Woo 2018; Sunmoo et al. 2022). Task 3 of the SMM4H workshop presented a benchmark classifier that achieves 0.96 F1-score by finetuning a pretrained BERTweet model (Klein et al. 2025, 2022). While this benchmark shows strong performance, our work aims to further explore the potential of BERT-based models and recent advancements in NLP, particularly Large Language Models (LLMs), to enhance detection accuracy in this domain. This paper presents our approaches for classifying tweets from dementia family caregivers on Twitter.



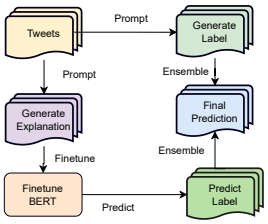


Figure 1: Overview of our approach.

## 2 System Description

Our system comprises two main components: a Bidirectional Encoder Representations from Transformers (BERT)-based module (Devlin et al. 2019) and a Large Language Model (LLM)-based module (Chang et al. 2024). The BERT-based module includes the BERT base (uncased) model (Devlin et al. 2019) and BERTweet pretrained on Twitter data (Nguyen, Vu, and Nguyen 2020). The LLM-based module uses the LLAMA-3 8B model (Grattafiori et al. 2024), integrated via the Unsloth framework (Daniel Han and team 2023) for efficient inference.

### 2.1 Problem and Dataset

Task 3 of the SMM4H Workshop is a binary classification task focused on identifying English tweets where users mention having a family member with dementia (Klein et al. 2025). The dataset includes 6,724 training, 353 validation, and 10,000 test tweets. Tweets are labeled 1 if the user reports a family member with dementia, and 0 if dementia is merely mentioned. The goal is to enable internet-based interventions using Twitter as a platform. Performance is evaluated using the $F_1$-score for class 1. A BERTweet-large baseline achieved an F1-score of 0.96 (Klein et al. 2022). Descriptive statistics are provided in Table 3 in the Appendix.

### 2.2 Model Structure

We implemented two ensemble methods for this task: the BERT Ensemble and the LLM Ensemble. The BERT Ensemble combines predictions from three models using majority voting: (1) a GloVe Twitter embedding-based model with Convolutional Neural Networks (CNN) (O'shea and

Nash 2015) and Long Short-Term Memory (LSTM) networks (Gers, Schmidhuber, and Cummins 2000), (2) BERT-base (uncased), and (3) BERTweet-base. Each model takes original tweets as input independently before aggregation. Ensemble learning effectively improves predictive performance and mitigates limitations of individual models, especially in complex tasks (Dong et al. 2020; Polikar 2012).

In contrast, the LLM Ensemble leverages explanation-based prompting and combines outputs from three components: (4) a few-shot prompted large language model (LLM), (5) BERT-base, and (6) BERTweet-base, both trained on the concatenation of the original tweet and a generated explanation. Explanations are produced via prompting and appended to tweets using a `[SEP]` token. Final predictions are determined through majority voting across the three sources. This explanation-augmented approach improves interpretability while maintaining strong classification performance. The prompts used for zero-shot explanation generation are listed in Table 5 (Appendix), and Table 4 provides an example explanation and a comparison between LLM-generated and BERT-based predictions.

### 2.3 Implementation Details

For the BERT-based models (BERT-base and BERTweet), experiments were conducted with 16GB of RAM and an NVIDIA RTX 3050 GPU with 6GB of memory. The maximum input sequence length was set to 128 tokens. We used the AdamW optimizer (Loshchilov, Hutter et al. 2017) with a learning rate of $2e-5$, and the Cross Entropy Loss function (Zhang and Sabuncu 2018) to update model weights.

The third model combined GloVe Twitter embeddings (Pennington, Socher, and Manning 2014) with a CNN and LSTM network. This model used the Adam optimizer (Zhang 2018) with a learning rate of 0.001 and Binary Cross Entropy loss (Zhang and Sabuncu 2018). The architecture included a hidden dimension of 100, 128 convolutional filters with a filter size of 5, and a dropout rate of 0.5.

A batch size of 15 was used for training all three models. For the ensemble approach, majority voting was applied across the predictions of all three models.

We used LLAMA-3.1 to generate both labels and explanations, with temperature set to 0.1 to reduce output confabulation. The instructions and sample examples used for few-shot prompting are provided in Table 5 in the Appendix. We selected 10 examples for the few-shot prompts and set the maximum sequence length to 1024 tokens. For zero-shot prompting, the maximum sequence length was set to 512 tokens. We used an A100 GPU with 40 GB of memory.

## 3 Evaluation

As shown in Table 1, training the BERT and BERTweet models on both the original tweets and their corresponding LLM-generated explanations yields slightly better performance compared to training solely on the original tweets. The results presented in this table are based on models trained using the provided training set and evaluated on the official validation set. Since the ground truth labels for the 10,000 test instances were not released, we report the test performance based on the official Codalab leaderboard, as summarized in Table 2. In particular, the proposed ensemble approach outperforms the individual models, achieving a modest improvement of 0.06 in the $F_1$-score.

| Approach | Precision | Recall | $F_1$-score |
|---|---|---|---|
| 1. GloVe Twitter + LSTM-CNN | 0.8917 | 0.9145 | 0.9030 |
| 2. BERT base (uncased) | 0.9322 | 0.9402 | 0.9362 |
| 3. BERTweet base | 0.9241 | 0.9359 | 0.9299 |
| 4. Few-shot prompting | 0.8565 | 0.7906 | 0.8222 |
| 5. BERT base on (Tweet + Explanation) | 0.9280 | 0.9359 | 0.9319 |
| 6. BERTweet base on (Tweet + Explanation) | 0.9218 | **0.9573** | 0.9392 |
| 7. **LLM Ensemble** (4, 5 and 6) | **0.9402** | 0.9402 | **0.9402** |

Table 1: Comparison of model performance across different approaches on the validation set

| Method | $F_1$-score | Precision | Recall |
|---|---|---|---|
| **LLM Ensemble** | **0.957** | 0.951 | **0.962** |
| **BERT Ensemble** | 0.951 | **0.961** | 0.941 |

Table 2: Performance of our submitted models on the test set

## 4 Discussion

Our results show that LLMs can generate informative explanations that offer valuable contextual insights beyond the original text. While the generated labels occasionally differed from human annotations, the accompanying explanations still enhanced the performance of both BERT-base and BERTweet models.

Minimizing confabulations in LLM outputs is crucial, particularly in healthcare contexts. To promote more accurate responses, we set the generation temperature to 0.1. Prompt design also plays a key role in shaping LLM behavior, and more targeted instructions are needed to improve output quality—an area for future exploration. Our error analysis revealed ambiguity in defining a close family member. For example, one tweet mentioned an uncle with dementia but was annotated as not referring to a family member, highlighting a need for clearer annotation guidelines. Lastly, the GloVe-based approach remained strong, demonstrating its value as a fast, resource-efficient alternative to more compute-intensive LLM methods.

## 5 Conclusion

LLMs like LLAMA show promise in detecting users with family members affected by dementia, but further work is needed. Due to resource constraints, we used only the 8B version of LLAMA-3, limiting larger-scale testing. Future work includes experimenting with prompts, testing models like Gemma-3 and Phi, using lightweight ensembles to reduce compute, and exploring fine-tuning and decoding strategies (e.g., temperature, top_p, top_k) to enhance performance and coherence.

## References


Alzheimer's Disease International. 2015. ADI - Dementia statistics — alzint.org. https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/. [Accessed 19-04-2025].

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3): 1–45.

Cheng, T. Y.-M.; Liu, L.; and Woo, B. K. 2018. Analyzing Twitter as a platform for Alzheimer-related dementia awareness: thematic analyses of tweets. *JMIR aging*, 1(2): e11542.

Daniel Han, M. H.; and team, U. 2023. Unsloth.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; and Ma, Q. 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14: 241–258.

Gers, F. A.; Schmidhuber, J.; and Cummins, F. 2000. Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10): 2451–2471.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Gryboski, L.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Mazzotti, D.; Onishi, T.; Powell, J.; Raithel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.

Klein, A. Z.; Magge, A.; O'Connor, K.; and Gonzalez-Hernandez, G. 2022. Automatically Identifying Twitter Users for Interventions to Support Dementia Family Caregivers: Annotated Data Set and Benchmark Classification Models. *JMIR Aging*, 5(3): e39547.

Loshchilov, I.; Hutter, F.; et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5: 5.

Moorhead, S. A.; Hazlett, D. E.; Harrison, L.; Carroll, J. K.; Irwin, A.; and Hoving, C. 2013. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, 15(4): e1933.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14.

O'shea, K.; and Nash, R. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Polikar, R. 2012. Ensemble learning. *Ensemble machine learning: Methods and applications*, 1–34.

Qiu, J.; Li, L.; Sun, J.; Peng, J.; Shi, P.; Zhang, R.; Dong, Y.; Lam, K.; Lo, F. P.-W.; Xiao, B.; et al. 2023. Large AI models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*, 27(12): 6074–6087.

Ravì, D.; Wong, C.; Deligianni, F.; Berthelot, M.; Andreu-Perez, J.; Lo, B.; and Yang, G.-Z. 2016. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1): 4–21.

Sunmoo, Y.; Broadwell, P.; Alcantara, C.; Davis, N.; Haeyoung, L.; Bristol, A.; Tipiani, D.; Nho, J. Y.; and Mittelman, M. 2022. Analyzing topics and sentiments from twitter to gain insights to refine interventions for family caregivers of persons with Alzheimer's disease and related dementias (ADRD) during COVID-19 pandemic. *Studies in health technology and informatics*, 289: 170.

Zhang, Z. 2018. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, 1–2. Ieee.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.


## 6 Appendix

| | Average token length | Average unique tokens | Max token length | Min token length | Number of documents | Number of positive labels |
|---|---|---|---|---|---|---|
| **Training** | 47.54 | 39.8 | 119 | 7 | 6724 | 4523 |
| **Valid** | 46.6 | 38.9 | 82 | 11 | 353 | 234 |
| **Test** | 47.8 | 40.01 | 119 | 6 | 10000 | NA |

Table 3: Descriptive statistics of the training, valid and test set

| Original Tweet | Orginal Label | LLM Generated Explanation | LLM Generated Label | Prediction after BERT on Explanation |
|---|---|---|---|---|
| Leeza Gibbons shares how her organization is responding to burdens caregivers have been feeling, and the lessons she learned from her late mom who had Alzheimer's. | 0 | The tweet explicitly states that Leeza Gibbons' mother had Alzheimer's, which aligns with the classification criteria for a close family member. | 1 | 0 |
| School horses that students ride are exactly like someone with dementia. I mean, I farm my mother with Alzheimers out to birthday parties and everything. | 1 | The tweet uses a metaphor to compare horses with dementia, but it does not explicitly state that the user has a family member with dementia. The phrase "farm my mother with Alzheimers" is used in a figurative sense, implying that the user takes care of their mother, but it does not confirm a family member with dementia. | 0 | 1 |
| Ambassador Yael Lempert:My American daughter & her husband live in London & trying to get an emergency passport for their baby daughter to return to USA - 90 year old grandmother failing due to Alzheimers. No one is helping them. Auto email responses & recorded messages. | 1 | The tweet refers to a 90-year-old grandmother failing due to Alzheimer's. However, it does not indicate that the user has a family member with dementia.It is a third-person narrative, which violates Rule 4. | 0 | 1 |
| Bunny girl senpai is too overrated. Mc's a pervert, his dream girl ain't real, and his little sister has dementia. Why do people like this? | 0 | The tweet explicitly states that the little sister has dementia, which aligns with the classification criteria for a close family member. | 1 | 0 |

Table 4: Comparison of the original tweet and label, LLM-generated explanation and label, and the final BERT prediction incorporating the LLM-generated explanation

| **Zero-shot Prompt Instruction** |
|---|
| You are an advanced language model trained to distinguish tweets that indicate having a family member with dementia (annotated as ”1”) from those that do not (annotated as ”0”).<br>For the purpose of this classification task, a family member is a close relative defined as any one of the following relationships to the Twitter user: parent, grandparent, great grandparent, sibling, spouse, or child, including any of these relevant relationships through marriage.<br>Output only ”1” if a tweet reports having a family member with dementia and ”0” otherwise. Generate an explanation also.<br><br>### Instruction:<br>Classify the following tweets as either ”1” or ”0”.<br>1. Tweets should be labeled as ”1” if they explicitely or implicitely indicate that the user has a family member with dementia.<br>2. Tweets should be labeled as ”0” if they self-report having dementia or refer to a family member beyond those defined as a close relative.<br>3. Tweets should be labeled as ”0” if they indicate knowing or being a caregiver for a person with dementia but do not specify a family member.<br>4. Tweets should be labeled as ”0” if they quote what was said by others.<br>5. Tweets should also be labeled as ”0” if they refer to a family member defined as a close relative but indicate that dementia is merely suspected.<br>6. Tweets should also be labeled as ’0’ if they mention dementia as a possibility but it’s not certain. |
| **Few-shot Prompt Examples** |
| # Tweet:<br>My 87 y/o grandfather with brain cancer got it, my 85 y/o grandmother with dementia got it, my 62 y/o mother with stage 4 OCP got it, my 59 y/o step-father with type 2 diabetes got it.Quit pretending like you know anything<br><br># Response:<br>Class: 1<br>Explanation: The tweet explicitly mentions the user’s grandmother, a close family member, has dementia. This meets Rule 1.<br><br># Tweet:<br>Sidenote on that, if someone tells u their parent has Alzheimer’s please suppress every desire to say your grandparent or great aunt did too. I appreciate the connection and that u can relate to the experience but it is so unbelievably different. Tell me a different time.<br><br># Response:<br>Class: 0<br>Explanation: This tweet quotes general advice and does not indicate that the user has a family member with dementia. It violates Rule 4 (quotes what was said by others or general reference). |

Table 5: Zero-shot and few-shot prompts used to generate the LLM labels and explanations.