

# BOUN at #SMM4H-HeaRD 2025: Enhancing Dementia Family Caregiver Detection on Twitter/X with a Lightweight Language Model

Ece Elif Adak, Şaziye Betül Özateş

Institute for Data Science and Artificial Intelligence,  
Boğaziçi University, İstanbul, Turkey  
{ece.adak@std.,saziye.ozates@}bogazici.edu.tr

## Abstract

Detecting caregivers of dementia family members on social media networks is becoming increasingly important for healthcare support. This paper presents our system for detecting dementia family caregivers on Twitter/X as part of the Social Media Mining for Health (#SMM4H) Shared Task 3. Using the Gemma 2-2B decoder-only transformer model with 2 billion parameters, we implemented 8-bit quantization and Low-Rank Adaptation (LoRA) to efficiently fine-tune on a dataset of 6,724 tweets. Our approach achieved an F1 score of 0.966 on the test set, marginally outperforming the previous state-of-the-art BERTweet-Large baseline and achieving the highest score in this task. The results suggest that larger, newer models as well as different model architectures can improve performance in this task, potentially opening avenues for better identification of caregivers, who may benefit from support resources on social media platforms.

## Introduction

The Social Media Mining for Health (#SMM4H) shared tasks address natural language processing (NLP), machine learning (ML), and artificial intelligence (AI) challenges inherent to utilizing social media data for health-related research. Task 3 is the detection of dementia family caregivers on Twitter/X. Many traditional methods were developed to try and provide support to family caregivers of people with dementia. Unfortunately, most of them are not in use and hence unable to provide support to people in need (Klein et al. 2025).

The age of social media networks has opened other avenues. People now use social media as a communication method, a news source, and as a way of expressing themselves. These qualities have allowed social media networks to become the foundation of many studies. Detection of dementia family caregivers is but one of them. This paper presents a system developed for this task using a dataset specifically designed to identify dementia family caregivers on Twitter/X (Klein et al. 2022) and reports the obtained results. We employ Gemma 2, a novel generalist model with a decoder-only architecture, achieving an F1 score of 0.966 in detecting dementia family caregivers within the tweet dataset provided by the #SMM4H Shared Task 3.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

## Related Work

Recent research has increasingly focused on leveraging AI methods such as machine learning, natural language processing, and large language models (LLMs) to identify family caregivers of people with dementia. This capability addresses a critical need, as early identification enables timely delivery of education, training, emotional support, and access to essential formal resources for caregivers (Turró-Garriga, Fernández-Adarve, and Monreal-Bosch 2022).

Social media platforms and online forums have emerged as valuable data sources for caregiver detection, serving as significant avenues for caregivers to share experiences and seek support. Researchers have successfully developed Machine Learning (ML) models to identify Twitter/X users who are family caregivers for people with dementia, achieving high accuracy rates. These approaches often employ supervised learning algorithms such as Support Vector Machines and Random Forests, known for their effectiveness in high-dimensional spaces and robustness (Tsoi et al. 2023).

Transformer models, particularly BERT (Devlin et al. 2019), have shown state-of-the-art performance in various NLP tasks related to caregiver identification. Beyond direct identification, NLP techniques are crucial for analyzing the content shared on these platforms. For example, NLP can identify individuals who self-disclose as dementia caregivers or frequently discuss caregiving topics (Tsoi et al. 2023). Furthermore, text mining, sentiment analysis, and topic modeling are used to understand caregivers' online communications; sentiment analysis can gauge the emotional tone in posts, offering insight into burden levels and support needs, while topic modeling helps identify prevalent themes and key challenges faced by caregivers, thereby revealing self-disclosure patterns and emotional states (Wang et al. 2021; Klein et al. 2022).

Electronic Health Records (EHRs) represent another significant data source. Rule-based NLP systems have been applied to clinical notes to identify mentions of family members and their caregiving roles (Tsoi et al. 2023). One study demonstrated high accuracy in identifying informal caregivers specifically from telephone encounter notes (Mahmoudi et al. 2022). Named entity recognition and relation extraction techniques can detect linguistic cues indicative of caregiving responsibilities, such as mentions of assisting with activities of daily living or managing medications

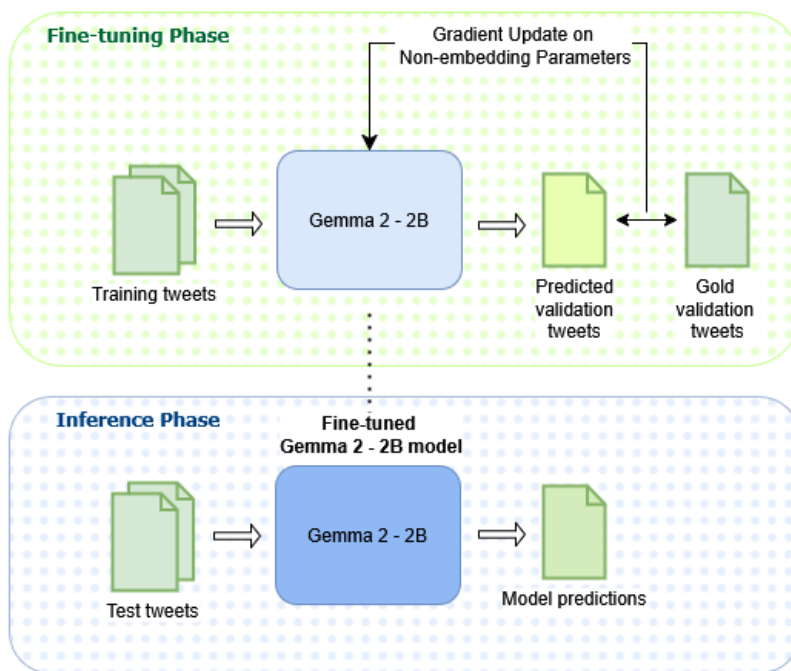


Figure 1: Fine-tuning and inference steps for the Gemma 2 - 2B model.

(Gilmore-Bykovskiy et al. 2019).

## Data

The dataset for Shared Task 3 consists of 6,724 tweets in the training set, 353 tweets in the validation set, and 1,769 tweets in the test set. Table 1 shows the label distribution in the entire dataset. Data was collected from publicly available tweets in the period of May 4 and May 21, 2021. Each selected tweet contains at least one dementia-related keyword from a selected list and language indicative of diagnosis. All tweets included in the dataset are in English, are not retweets and are posted by different users.

Dataset	Label 1	Label 0	Total
Training	4523	2201	6724
Validation	234	119	353
Test	unknown	unknown	1769

Table 1: Dataset composition of the #SMM4H-HeaRD Shared Task 3. Label 1 indicates the positive class (tweets related to dementia caregivers) and Label 0 indicates the negative class.

## Methodology

Transformer-based language models have demonstrated exceptional performance in text classification tasks, particularly when fine-tuned with task-specific data. The Gemma family of models, developed by Google, builds on these advancements with a lightweight yet powerful architecture designed for both generative and classification tasks. The

experiments documented in this paper utilized the Gemma 2 model with a parameter size of 2 billion (2B). As part of the Gemma family, the Gemma 2 models are built on a decoder-only transformer architecture. This architectural choice means that the model processes input sequences and generates output tokens autoregressively, making it well-suited for text generation tasks, while also being adaptable for text classification.

Specifically, the Gemma 2 - 2B variant comprises 590,118,912 embedding parameters, which handle the representation of input tokens, and 2,024,517,888 non-embedding parameters, which constitute the core transformer blocks responsible for computation and learning complex patterns. The model was trained primarily on a large corpus of data in English, influencing its linguistic capabilities and optimal use cases (Team et al. 2024).

## Training

We fine-tuned Gemma 2 on Google Colab using an A100 GPU with 40 GB RAM. To facilitate fine-tuning a model of this size on an A100, the BitsAndBytesConfig<sup>1</sup> class was used to set the model to 8-bit precision instead of the original 32-bit precision of the open weights. The training process was further optimized for speed by the application of parameter-efficient fine-tuning methods. Low-Rank Adaptation (LoRA) was used to select 20,771,328 parameters for fine-tuning (Hu et al. 2021; Xu et al. 2023). These implementations achieved a training speed comparable to base BERT models. Figure 1 depicts the the training scheme of

<sup>1</sup><https://huggingface.co/docs/transformers/en/quantization/bitsandbytes>

our model.

The model was trained on the 6,724 tweets in the training set using a batch size of 8 for a maximum of 5 epochs. Early stopping was employed with the F1 score as the evaluation metric to determine the best model. Monitoring the training and validation losses indicated that 5 epochs or fewer were sufficient to effectively train the Gemma 2 – 2B model under these conditions. Table 2 presents the hyperparameter settings used for training.

Hyperparameter	Value
Batch size	8
Gradient Accumulation Steps	4
Learning rate	2e-4
Learning rate decay	yes
Learning rate scheduler	linear
Weight decay	0.01
Optimizer	AdamW
Epochs	max. 5
Early stop	patience 500 steps
Precision	8-bit precision
Max. sequence length	128
<b>LoRA Hyperparameters</b>	
LoRA Rank	16
LoRA Alpha	32
Dropout Rate	0.1

Table 2: Hyperparameters of our model during the fine-tuning phase.

## Results

The original study that produced the dataset used in this work tested the performance of several classifiers, identifying BERTweet-Large as the top performer at the time with an F1-score of 0.962 on the test set (Klein et al. 2022).

BERTweet-Large is a highly specialized model focused on English Twitter/X data. It is based on the BERT architecture, meaning it is an encoder-only transformer model with approximately 355 million parameters – significantly smaller than Gemma 2 variants. The key innovation of the BERTweet model lies in its pre-training strategy and data. It was pre-trained from scratch using the RoBERTa pre-training procedure (which optimizes BERT’s training) on an enormous corpus of 850 million English Tweets (80GB). This domain-specific pre-training makes BERTweet exceptionally effective at understanding the nuances of social media language; including slang, hashtags, mentions, emojis, and the generally noisy and informal nature of tweets. BERTweet demonstrated state-of-the-art performance (at the time of its release) specifically on tweet-related NLP tasks such as part-of-speech tagging, named entity recognition, and text classification within the Twitter/X domain. A cluster of 8 A100 GPUs with 40 GB RAM was used during pre-training (Nguyen, Vu, and Tuan Nguyen 2020).

Table 3 presents a comparison between the models in the benchmark study (Klein et al. 2022) and our approach on the test set. The results reveal the superior performance of our

Classifier	Prec.	Recall	F1-score
SVM	0.884	0.939	0.910
BERT-Base-Uncased	0.924	0.954	0.938
DistilBERT-Base-Uncased	0.930	0.942	0.936
RoBERTa-Large	0.918	<b>0.982</b>	0.949
BioBERT-Large-Cased	0.907	0.978	0.941
Bio+ClinicalBERT	0.903	0.958	0.930
BERTweet-Large (baseline)	0.946	0.979	0.962
Gemma 2-2b (our model)	<b>0.957</b>	0.976	<b>0.966</b>

Table 3: Classifier performance of our model and the models presented in the benchmark study (Klein et al. 2022) on the test set.

model compared to all other models, achieving an F1 score of 0.966, which is the best obtained in this shared task.

According to the precision metric, we observe that our Gemma-based model surpasses the strong baseline BERTweet-Large model by a considerable margin in precision, achieving a 1.1% higher score. This indicates that when our model predicts the positive class, it is considerably more likely to be correct, resulting in fewer false positives compared to the baseline. While our model shows this superior precision, it exhibits a marginal decrease in recall by 0.3% compared to the baseline, meaning it fails to identify a slightly larger fraction of the true positive instances. However, given that the F1 score is the primary evaluation metric for this task, the substantial gain in precision more than offsets the slight reduction in recall. This results in a higher overall F1 score, demonstrating that our model achieves a more effective balance between precision and recall for this specific classification challenge compared to the baseline model.

Classifier	Precision	Recall	F1-score
submission 1	0.954	0.974	0.964
submission 2	0.942	0.97	0.956

Table 4: Performance of our fine-tuned Gemma 2 - 2B model on the validation set.

Classifier	Precision	Recall	F1-score
submission 1	0.957	0.976	0.966
submission 2	0.958	0.968	0.963

Table 5: Performance of our fine-tuned Gemma 2 - 2B model on the test set.

A comparison of the results in Table 4 and Table 5 reveals that the performance of our model is slightly better on the test set than on the validation set. We attribute this, at least in part, to the substantial size discrepancy between the two datasets — 353 samples in the validation set versus 1,769 in the test set. The larger size of the test set likely contributes to more reliable and accurate evaluation, leading to the observed performance improvement.

## Discussion

The experimental results demonstrate that further improvements are possible with newer approaches by achieving an improved F1 score of 0.966 with a Gemma 2 variant, a model from a family of generalist models. It is important to consider that due to the need to use LoRA to optimize the training process only 20 million parameters were fine-tuned.

We observe that a generic model like Gemma 2 - 2B can outperform the domain-specific BERTweet-Large, which was specifically trained for the Twitter/X domain. This finding suggests that larger, more general models could potentially eliminate the need for models tailored to specific domains. However, relying solely on larger models is not necessarily advantageous. The environmental cost of training and deploying such models is substantial, as they consume significantly more computational power and energy. This raises questions about the trade-off between marginal performance gains and the growing carbon footprint associated with increasingly large models. Exploring more efficient fine-tuning methods or smaller, task-specific models may offer a more sustainable path forward.

## Conclusion

We selected a powerful and advanced model for fine-tuning to detect dementia family caregivers on Twitter/X. The methods employed ensured that the training could be carried out on a single A100 GPU while maintaining sufficient speed to allow for experimentation. The model trained for this task, Gemma 2 - 2B, marginally outperformed the strong baseline model. The experimental results indicate that, although the previous state-of-the-art model performs reliably and shows excellent results, performance improvements are still possible, even if gradual. Several factors may contribute to this, including the relative newness of the Gemma 2 models, which incorporate architectural advancements learned from the Gemini models, as well as its larger size (Team et al. 2024). Additionally, the model's architecture may have played a role. In the benchmark study (Klein et al. 2022), all classifiers except the Support Vector Machine (SVM) were BERT-based, meaning they share an encoder-only architecture. In contrast, Gemma 2 models are decoder-only, and the 2B variant used in this study is the smallest in the model family. Other decoder-only models or encoder-decoder models could also outperform the baseline model.

The performance of the models trained for this task can still be improved. The previously mentioned class imbalance can be improved by the use of synthetic data. State-of-the-art LLMs such as GPT-4.1, Gemini 2.5 or Claude 3.7 can be used to generate synthetic tweets similar to the ones in the training set. Further experimentation with Gemma 2 and different LoRA-derived methods can yield even better results.

## References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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. Minneapolis, Minnesota: Association for Computational Linguistics.

Gilmore-Bykovskiy, A.; Mullen, S.; Block, L.; Jacobs, A.; and Werner, N. E. 2019. Nomenclature Used by Family Caregivers to Describe and Characterize Neuropsychiatric Symptoms. *The Gerontologist*, 60(5): 896–904.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

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.

Mahmoudi, E.; Wu, W.; Najarian, C.; Aikens, J.; Bynum, J.; and Vydiswaran, V. G. V. 2022. Identifying Caregiver Availability Using Medical Notes With Rule-Based Natural Language Processing: Retrospective Cohort Study. *JMIR Aging*, 5(3): e40241.

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

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; Ferret, J.; Liu, P.; Tafti, P.; Friesen, A.; Casbon, M.; Ramos, S.; Kumar, R.; Lan, C. L.; Jerome, S.; Tsitsulin, A.; Vieillard, N.; Stanczyk, P.; Girgin, S.; Momchev, N.; Hoffman, M.; Thakoor, S.; Grill, J.-B.; Neyshabur, B.; Bachem, O.; Walton, A.; Severyn, A.; Parrish, A.; Ahmad, A.; Hutchison, A.; Abdagic, A.; Carl, A.; Shen, A.; Brock, A.; Coenen, A.; Laforge, A.; Paterson, A.; Bastian, B.; Piot, B.; Wu, B.; Royal, B.; Chen, C.; Kumar, C.; Perry, C.; Welty, C.; Choquette-Choo, C. A.; Sinopalnikov, D.; Weinberger, D.; Vijaykumar, D.; Rogozińska, D.; Herbison, D.; Bandy, E.; Wang, E.; Noland, E.; Moreira, E.; Senter, E.; Eltyshv, E.; Visin, F.; Rasskin, G.; Wei, G.; Cameron, G.; Martins, G.; Hashemi, H.; Klimczak-Plucińska, H.; Batra, H.; Dhand, H.; Nardini, I.; Mein, J.; Zhou, J.; Svensson, J.; Stanway, J.; Chan, J.; Zhou, J. P.; Carrasqueira, J.; Iljazi, J.; Becker, J.; Fernandez, J.; van Amersfoort, J.; Gordon, J.; Lipschultz, J.; Newlan, J.; yeong Ji, J.; Mohamed, K.; Badola, K.; Black, K.; Millican, K.; McDonnell, K.; Nguyen, K.; Sodhia, K.; Greene, K.; Sjoesund, L. L.; Usui, L.; Sifre, L.; Heuermann,

L.; Lago, L.; McNealus, L.; Soares, L. B.; Kilpatrick, L.; Dixon, L.; Martins, L.; Reid, M.; Singh, M.; Iverson, M.; Görner, M.; Velloso, M.; Wirth, M.; Davidow, M.; Miller, M.; Rahtz, M.; Watson, M.; Risdal, M.; Kazemi, M.; Moynihan, M.; Zhang, M.; Kahng, M.; Park, M.; Rahman, M.; Khatwani, M.; Dao, N.; Bardoliwalla, N.; Devanathan, N.; Dumai, N.; Chauhan, N.; Wahltinez, O.; Botarda, P.; Barnes, P.; Barham, P.; Michel, P.; Jin, P.; Georgiev, P.; Culliton, P.; Kuppala, P.; Comanescu, R.; Merhej, R.; Jana, R.; Rokni, R. A.; Agarwal, R.; Mullins, R.; Saadat, S.; Carthy, S. M.; Cogan, S.; Perrin, S.; Arnold, S. M. R.; Krause, S.; Dai, S.; Garg, S.; Sheth, S.; Ronstrom, S.; Chan, S.; Jordan, T.; Yu, T.; Eccles, T.; Hennigan, T.; Kocisky, T.; Doshi, T.; Jain, V.; Yadav, V.; Meshram, V.; Dharmadhikari, V.; Barkley, W.; Wei, W.; Ye, W.; Han, W.; Kwon, W.; Xu, X.; Shen, Z.; Gong, Z.; Wei, Z.; Cotruta, V.; Kirk, P.; Rao, A.; Giang, M.; Peran, L.; Warkentin, T.; Collins, E.; Barral, J.; Ghahramani, Z.; Hadsell, R.; Sculley, D.; Banks, J.; Dragan, A.; Petrov, S.; Vinyals, O.; Dean, J.; Hassabis, D.; Kavukcuoglu, K.; Farabet, C.; Buchatskaya, E.; Borgeaud, S.; Fiedel, N.; Joulin, A.; Kenealy, K.; Dadashi, R.; and Andreev, A. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.

Tsoi, K. K. F.; Jia, P.; Dowling, N. M.; Titiner, J. R.; Wagner, M.; Capuano, A. W.; and Donohue, M. C. 2023. Applications of artificial intelligence in dementia research. *Cambridge Prisms: Precision Medicine*, 1: e9.

Turró-Garriga, O.; Fernández-Adarve, M. d. M.; and Monreal-Bosch, P. 2022. Needs Detection for Carers of Family Members with Dementia. *Healthcare*, 10(1).

Wang, Z.; Zou, N.; Xie, B.; Luo, Z.; He, D.; Hilsabeck, R.; and Aguirre, A. 2021. *Characterizing Dementia Caregivers' Information Exchange on Social Media: Exploring an Expert-Machine Co-development Process*, 47–67. ISBN 978-3-030-71291-4.

Xu, L.; Xie, H.; Qin, S.-Z. J.; Tao, X.; and Wang, F. L. 2023. Parameter-Efficient Fine-Tuning Methods for Pre-trained Language Models: A Critical Review and Assessment. arXiv:2312.12148.