Automatic Depression Detection Using An Interpretable Audio-textual Multi-modal Transformer-based Model

Mehrshad Saadatinia
saadatin@usc.edu

Pin-Tzu Lee
pintzule@usc.edu

Sreya Reddy Chinthala
chinthal@usc.edu

Om Jodhpurkar
jodhpurk@usc.edu

Sneh Thorat
snehpram@usc.edu

Abstract

In this study, we propose a multi-modal transformer-based framework for the detection of depression using audio and text modalities from clinical interview data. Our approach leverages the self-attention mechanism in transformers and improves both diagnostic accuracy and interpretability. The proposed model achieves an improvement in diagnostic accuracy for audio-textual depression detection in two benchmark datasets. Furthermore, by analyzing attention weights, we identify the most influential features in audio and text that drive the model's predictions, offering valuable insights into key factors contributing to depression diagnosis.

1 Introduction

Depression is a common mental disorder marked by persistent low mood, loss of interest, and fatigue, and in severe cases, can lead to suicide. Despite its impact, depression is often under-diagnosed and under-treated due to costly, time-consuming diagnostics and patients' reluctance to disclose symptoms for fear of stigma. This highlights the need for automated depression detection systems that offer private assessments, encouraging more people to seek help.

This study aims to enhance both the accuracy and interpretability of automatic depression detection. By leveraging a multi-modal transformer-based approach, we aim to facilitate the diagnostic process and build trust in deep learning methods. Specifically, we investigate whether a multi-modal transformer model can improve the accuracy and interpretability of depression detection in clinical settings and how its performance compares to other models for this task.

1.1 Contributions

In this study, we propose an interpretable multi-modal transformer-based framework for automated depression detection using audio and text modalities derived from clinical interview data. Our contributions are summarized as follows:

1. We design a novel transformer-based architecture that integrates audio and text modalities for depression detection. By leveraging the self-attention mechanism, our model captures cross-modal interactions, enabling more comprehensive representations of clinical data.

2. The proposed approach achieves a significant improvement in performance in diagnostic accuracy compared to state-of-the-art methods across two benchmark datasets.

3. To enhance the transparency of the model, we analyze the attention weights. This allows us to identify the most influential features in the input that contribute to the model's predictions.

These contributions collectively advance the state-of-the-art in multi-modal depression detection by improving accuracy, interpretability, and clinical applicability, paving the way for more reliable and transparent automated diagnostic systems.

2 Related Work

The detection of depression has been widely explored using traditional machine learning approaches. However, for our study, we focus on related work that has employed deep learning. Many earlier works have utilized RNN-based networks, more commonly LSTMs, for detecting depression using either audio, text, or both Amanat et al. (2022). The proposal of the Transformer architecture by Vaswani et al. (2017) revolutionized NLP and approaches to tasks involving sequential data. Transformer architectures outperform traditional models such as CNNs and LSTMs by capturing long-range dependencies in both text and audio

features. Devlin et al. (2019) demonstrated the effectiveness of Transformer models, specifically BERT, for various NLP tasks, inspiring the use of Transformers in depression detection tasks as well.

Ye et al. (2021) utilized a multi-modal model combining text and audio, employing a Transformer for the text modality and a TCN for audio data; however, the potential of Transformers in audio remains under-explored in their study. In the work of Xiao et al. (2021), a BERT encoder is applied to text data, while an attention-based CCNN encodes audio features, followed by a fully connected joint encoding layer. Similarly, Guo et al. (2022) explored the use of RoBERTa for the text modality (Liu et al., 2019) and CNN-based encoding for audio, achieving state-of-the-art results on the DAIC-WOZ dataset, which is also used in our study.

While these approaches are promising, they lack interpretability. One of the most notable related works is the straightforward detection method proposed by Shen et al. (2022), which incorporates ELMo embeddings for text (Peters et al., 2018) and NetVLAD embeddings for audio (Arandjelovic et al., 2016). This method serves as a baseline for comparison in our work. Furthermore, few studies emphasize the explainability and interpretability of these models, a gap our project aims to address by employing XAI and interpretable methods.

3 Methodology

Our methodology primarily involves transformer-based encoding, fusion of two modalities, and classification of the joint embedding using a fully connected network. For the text modality, we utilize BERT for tokenization and embeddings. For the audio modality, we encode the entire audio track for each data instance into a 128-dimensional vector using a CNN-based encoder (NetVLAD). To enhance interpretability, we analyze the attention weights of the text modality to identify the most attended tokens, providing insights into the model's decision-making process.

3.1 Datasets

For this study, we utilize two primary datasets: the DAIC-WOZ (Distress Analysis Interview Corpus Wizard-of-Oz) dataset (Gratch et al., 2014) and the EATD (Emotional Audio-Textual Dataset) Corpus (Shen et al., 2022).

The DAIC-WOZ dataset comprises clinical interviews designed to aid in diagnosing psychological distress, including depression, anxiety, and post-traumatic stress disorder (PTSD). It contains both audio and text data, with annotated features such as linguistic signals, speech prosody, and depression labels based on PHQ-8 scores. The recordings are extensive, often lasting several minutes, making the dataset relatively large.

The EATD Corpus consists of Chinese audio and textual interview data annotated with depression levels based on SDS (Self-Rating Depression Scale) scores. Unlike DAIC-WOZ, the EATD Corpus has shorter recordings, typically less than a minute, and is publicly accessible. The dataset also includes preprocessed features such as alignment and noise removal, making it easier to use for experimentation.

Table 1 summarizes the key characteristics of the two datasets, including their depression metrics, participant counts, and recording lengths.

Name	Metric	Participants	Language	Size
DAIC-WOZ	PHQ-8	142 (42 depressed)	English	86GB
EATD	SDS	162 (30 depressed)	Chinese	740MB

Table 1: Dataset Details Used in This Study

4 Processing and Feature Extraction

4.1 DAIC-WOZ Dataset

The DAIC-WOZ dataset consists of clinical interviews, with each participant having a single text transcript and a corresponding raw interview audio file. Each transcript contains rows with start and end timestamps, speaker information (interviewer or interviewee), and spoken text. Preprocessing steps were applied to isolate the relevant information and prepare the data for feature extraction:

- **Text:** We extracted only the interviewee's transcripts and concatenated all their sentences into a single text.
- **Audio:** Using the timestamp metadata from the transcript file, we isolated and extracted audio segments corresponding to the interviewee.

4.2 EATD Dataset

The EATD dataset contains three recordings per subject, labeled as positive, negative, and neutral, along with their corresponding sentence transcripts. Each recording corresponds to a sentence expressing the respective sentiment. In addition to the

recordings, the dataset includes a label file for each subject, providing the SDS (Self-Rating Depression Scale) score. Subjects with scores above the standard threshold of 53 are classified as depressed.

Most of the necessary preprocessing steps, such as audio denoising, normalization, and modality alignment, have been performed by the dataset authors (Shen et al., 2022). However, we applied additional processing steps to prepare the dataset for our use:

- **Text:** The sentence transcripts are read from the files and encoded using a pre-trained embedding model. The token-level sentence embeddings are then concatenated to form a single representation for each subject.
- **Audio:** The audio signals are read from the `.wav` files and encoded using the NetVLAD model, which produces a 128-dimensional encoding vector for each recording. Similar to the text modality, we concatenate the three recording embeddings for each subject to obtain a single representation.

4.2.1 Data Augmentation

Since the dataset is heavily imbalanced we have used a data-augmentation method to artificially increase the number of depressed instances. We have created 6 different permutations of concatenation of the 3 embeddings (of audio and text) for each subject and effectively increased the number of positive instances 6-fold.

4.3 Text Feature Extraction

Text feature extraction is performed slightly differently for each dataset; however, the same underlying principles and overall methodology are applied to both datasets. For the EATD dataset, we utilize pretrained embeddings to encode the sentences prior to training, while contextualized embeddings are further learned using BERT during the training procedure. In contrast, for the DAIC-WOZ dataset, we tokenize the text using the BERT tokenizer, and the embeddings are learned entirely during training without the use of pretrained embeddings.

4.4 Audio Feature Extraction

We represent audio data using mel-spectrograms, a visual representation of sound that captures the amplitude of frequency components over time, preserving a high level of detail. The NetVLAD encoder layer (Arandjelovic et al., 2016) is applied to these spectrograms to generate compact embedded representations. This process yields 128-dimensional audio embeddings, which are subsequently concatenated. These audio encodings are then fed into a Transformer encoder to extract contextualized representations, which are used for downstream classification.

5 Implementation

To implement our proposed model, we used NetVLAD and BERT embeddings for audio and text modalities, respectively, as inputs to the Transformer Encoders. For each modality, the Transformer Encoder consists of a single layer with 4 attention heads. The outputs from the audio and text branches are concatenated to form a joint embedding. This joint embedding is then passed through a ReLU activation layer, followed by a fully connected (FC) layer of size 128 for additional feature extraction. Finally, the extracted features are passed through a classification layer to determine whether the subject is depressed.

6 Experiments and Results

6.1 EATD

We trained the transformer-based multi-modal network for 200 epochs and used 3-fold cross-validation for training and evaluation (one-third of the dataset was used for evaluation, and the rest was used for training each time). A learning rate of 0.0001 was used along with the Adam optimizer with mini-batches of size 32.

6.1.1 Evaluation

We implemented an LSTM-based version of our model to demonstrate the effectiveness and superiority of the transformer-based approach. To further emphasize the necessity of multi-modal methods, we implemented single-modal versions of both LSTM-based and transformer-based approaches. Additionally, we compared our results with the state-of-the-art findings reported by Shen et al. (2022) on the EATD dataset to provide a more comprehensive evaluation of our model's performance. The same parameter settings were applied to train all baseline models for a fair comparison.

As shown in Table 2, our proposed transformer-based method achieves an F1 score of **0.82**, outperforming all baseline models.

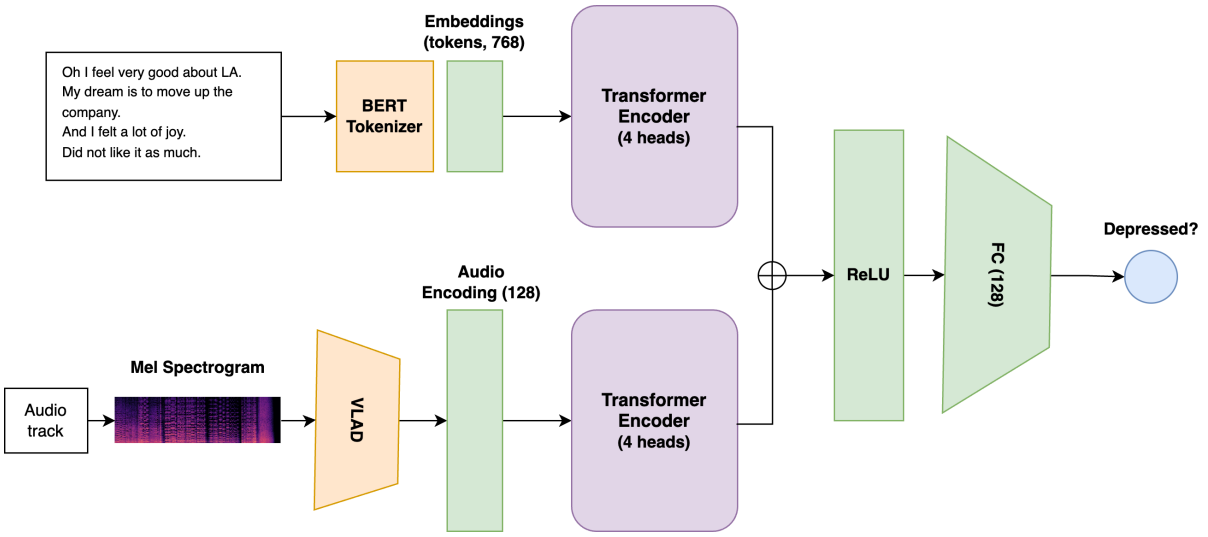

Figure 1: Schematics of the Proposed Methodology

Modality	Models	F1 Score
Audio	LSTM	0.72
Text	LSTM	0.79
Audio	Transformer	0.79
Text	Transformer	0.80
Text & Audio	Bi-LSTM & GRU + Attention (Shen et al., 2022)	0.71
Text & Audio	**Transformer (Proposed)**	**0.81**

Table 2: Comparison of Experimental Results on the EATD Corpus Between Our Method and the Baseline Models Based on F1 Score

Modality	Models	Accuracy (%)
Audio	Transformer	72.0
Text	Transformer	75.8
Text & Audio	Bi-LSTM & GRU + Attention (Shen et al., 2022)	73.1
Text & Audio	Topic-Attentive Transformer (Guo et al., 2022)	73.9
Text & Audio	**Transformer (Proposed)**	**75.8**

Table 3: Comparison of Experimental Results on the DAIC-WOZ Corpus Between Our Method and the Baseline Models Based on Accuracy

6.2 DAIC-WOZ

We trained the transformer-based multi-modal network on the DAIC-WOZ dataset for 40 epochs. Since this corpus contains much longer sentences, we standardized all sentence lengths to 50 tokens. Longer sentences were truncated, and shorter ones were padded to ensure consistency. The network was trained using a batch size of 8 and the AdamW optimizer with a learning rate of 4×10^{-5}. After convergence, the training accuracy reached 100%, while the test accuracy achieved **75.8%**.

6.2.1 Evaluation

We implemented single-modal versions of the architecture for the audio and text modalities separately to provide a comparative analysis of the effectiveness of the multi-modal approach. Additionally, we included state-of-the-art results from previous studies to contextualize our findings and demonstrate the empirical advantages of our method. Table 3 summarizes the results of this comparison.

7 Interpretability: Attention Visualization

To better interpret the decision-making process of our transformer model, we visualized the self-attention weights for the text modality. Specifically, we extracted the matrix of keys against queries, where the magnitude of each element represents the amount of attention each query pays to different keys. In the self-attention mechanism, keys and queries are derived from the same sentence; therefore, the rows and columns of this matrix correspond to the same tokens. These matrices can be visualized using heatmap plots, which help us better understand the alignment of attention in key-query pairs and the model's decision-making process. This visualization can also serve to verify whether the model is functioning as expected.

Another advantage of this method is that it allows us to approximate the role of each attention head. For example, one attention head might focus on emotion-related information, while another may emphasize location-like details. Although there are no exact mechanisms to fully interpret what each head is doing, this approach provides an approximate understanding of the relevance of each

head.

In Figure 2, we provide heatmaps from Head 0 of the self-attention mechanism for one of the sentences. The resulting heatmap offers insight into the focus of the model across the input tokens.

- Localized attention patterns are observed between tokens such as *"smart," "very,"* and *"uh,"* which likely reflect meaningful semantic relationships related to emotions and their expression within the sentence.
- The attention weights reveal localized interactions between semantically related tokens, highlighting the model's ability to capture short-range dependencies that are essential for understanding contextual patterns in text.
- Padding tokens exhibit near-zero attention, demonstrating the model's ability to ignore non-informative inputs and focus solely on meaningful content.

The observed attention weights suggest that the model focuses on relevant linguistic cues, such as emotionally charged adjectives, adverbs, and filler words (e.g., *"smart," "very," "uh"*), which are significant features in depression detection tasks. This visualization highlights the role of attention mechanisms in identifying key tokens that contribute to the classification outcome, aligning with the objectives of Explainable AI (XAI) by providing transparency and interpretability.

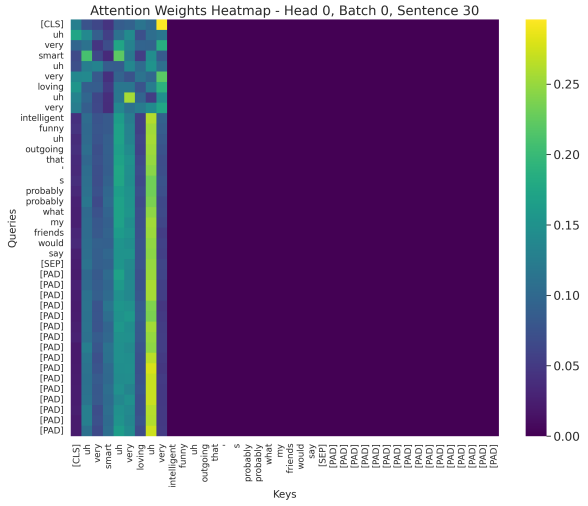

Figure 2: Attention weight visualization from the transformer model, showing strong focus on self-referential tokens and activity-related words, while effectively ignoring padding tokens.

Additionally, we plotted heatmaps for all 4 attention heads of a layer for a single sentence to approximate the role of each head. Figure 3 shows the plots for the 4 attention heads. By examining these plots, we observe that each head attends to different entities while exhibiting some common patterns, such as ignoring irrelevant tokens. This confirms that our model is functioning as expected. Furthermore, the distinct focus of each attention head demonstrates that multi-headed attention effectively captures diverse aspects of the input, aligning with its intended purpose.

However, our current interpretability approach is limited to the text modality because the audio input is encoded into a single vector, which prevents us from extracting token-by-token attention weights. Addressing this limitation by developing methods to analyze fine-grained attention for audio features could be a promising direction for future research.

8 Conclusion

In this study, we proposed a multi-modal transformer-based model for automatic depression detection by integrating audio and text data. The proposed model demonstrated superior performance over traditional baselines and state-of-the-art methods, showcasing the strength of the transformer architecture in capturing complex relationships across multiple modalities. By leveraging transformers for both text and audio features, our approach effectively models contextual and semantic dependencies, which are critical for understanding nuanced patterns in depression-related data.

To improve the interpretability of the model, we visualized the self-attention weights, revealing that the model focuses on self-referential tokens, emotionally charged adjectives, and activity-related linguistic cues commonly associated with depressive tendencies. This analysis aligns with the principles of Explainable and interpretable AI and enhances transparency, helping to build trust in the model's decision-making process.

Our results emphasize the effectiveness of a multi-modal approach, where the fusion of text and audio features leads to significant improvements in both accuracy and interpretability compared to single-modal models. Furthermore, the attention head analysis demonstrates the benefits of multi-headed attention in capturing diverse aspects of the input data.

In future work, we aim to further enhance model

performance by incorporating more advanced fusion techniques and optimizing the transformer architecture. We also plan to explore additional sources of behavioral and linguistic data to broaden the model's applicability to diverse real-world clinical settings, contributing to more reliable and interpretable depression detection systems. Furthermore, applying the same interpretability techniques to the audio modality to identify important acoustic cues presents another promising direction for future research.

Appendix

Supplementary Figures

References

Amna Amanat, Muhammad Rizwan, Abdul Rehman Javed, Maha Abdelhaq, Raed Alsaqour, Sharnil Pandya, and Mueen Uddin. 2022. Deep learning for depression detection from textual data. *Electronics*, 11(5):676.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Amir Nazarian, Stephen Wood, Jill Boberg, David DeVault, Stacy Marsella, and David Traum. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Yanrong Guo, Chenyang Zhu, Shijie Hao, and Richang Hong. 2022. A topic-attentive transformer-based model for multimodal depression detection. *arXiv preprint arXiv:2206.13256*.

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*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jing Xiao, Yongming Huang, Guobao Zhang, and Wei Liu. 2021. A deep learning method on audio and text sequences for automatic depression detection. In *2021 3rd International Conference on Applied Machine Learning (ICAML)*, pages 388–392. IEEE.

Jiayu Ye, Yanhong Yu, Qingxiang Wang, Wentao Li, Hu Liang, Yunshao Zheng, and Gang Fu. 2021. Multi-modal depression detection based on emotional audio and evaluation text. *Journal of Affective Disorders*, 295:904–913.

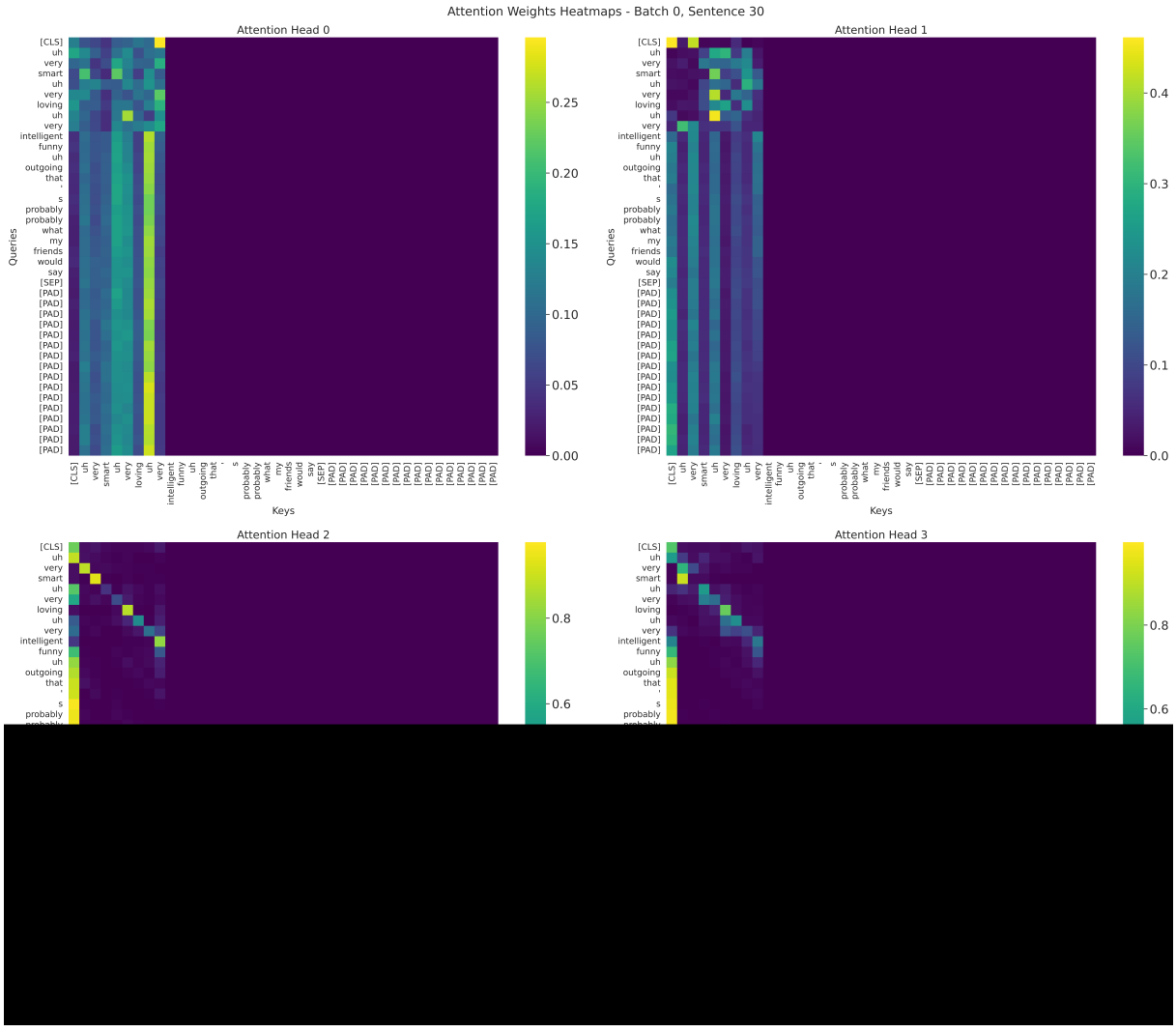

Figure 3: Attention weights for key-query pairs of all 4 heads of a layer in the transformer encoder.