

Course Title: Research Methodology in Computer Science

Course Code: AI800

Weighting: 20%

Lecturer: Dr. Mohammed Anbar

Hand-out Date: 25 Oct 2025

Hand-in Date: 08 Nov 2025

Hand-in Method: Online Submission via E-Learning.

Assignment Report: Literature Review Investigating Emotion –
Cognition Correlations in Transformer-Based Language Models: A
Neuropsychological Perspective.

Table of Contents

- Summary of Research papers
- Critical Analysis
- Connections
- Conflicts between the findings
- Synthesis Matrix for Literature Review
- References

Summary of Research Papers

Paper 1: Contextual Emotion Recognition Using Transformer-Based Models

The main objective of this work is to create a transformer-based system for context-aware emotion recognition, specifically utilizing the BERT paradigm. Its primary goal is to use textual context to improve the sensitivity and accuracy of emotion recognition. The study demonstrates how transformer models—particularly BERT and its offshoots—effectively capture intricate contextual clues, enhancing emotion identification capabilities. Because of its bidirectional training and pre-training on massive text corpora, BERT is better able to comprehend linguistic subtleties than one-directional models, which makes it appropriate for a variety of NLP applications. According to the article, such context-aware emotion models may result in AI interactions that are more sympathetic and human-like. Contextual emotion recognition experiments demonstrate that transformer models exhibit comparable patterns to human emotional cognition, which is also extremely context-dependent. However, because current research mostly prioritizes performance measurements above theoretical knowledge, it also highlights issues like bias in emotion recognition and the need for more thorough investigation of emotion–cognition relations (Devgan, 2023).

Paper 2: Representing Affect Information in Word Embeddings

By investigating whether the affective meaning of words—which is determined by valence, arousal, and dominance—is encoded in pre-trained word embeddings, the study seeks to understand how language models capture human-like emotional knowledge. The researchers performed correlational and classification analyses across various embedding types, varied in their contextualization and emphasis on emotion during training, using a human-labeled dataset as ground truth. The findings demonstrated that BERT embeddings could only successfully capture emotional information when they were refined on tasks involving emotions or exposed to situations that were emotionally charged. Although the work offers insightful information on how emotions are modeled in language, it lacks integration with contextual, cognitive, and multimodal aspects of emotion representation and is conceptually and methodologically broad (Zhang et al., 2022).

Paper 3: Enhancing Cognitive Model of Emotions with Representation Learning

In order to depict complex human emotions in accordance with psychological models such as Plutchik's emotion wheel, the paper presents transformer-based contextual embeddings. The authors create an emotion graph that reveals hierarchical and relational connections among

emotions by training a probing model to examine how emotions are encoded across deep learning layers. The paper demonstrates that similar emotions (such as joy and excitement) cluster tightly in embedding space by bridging computational representations with psychological emotion theory. The resulting emotion structures show that embedding spaces capture emotional relationships rather than considering emotions as discrete categories, and they meaningfully, if not precisely, correlate with human emotional models (Guo & Choi, 2021).

Paper 4: Emotion Aware: An Artificial Intelligence Framework for Adaptable, Robust, Explainable, and Multi-Granular Emotion Analysis

The study recommends the Emotion AWARE framework, which aims to enhance artificial intelligence systems' comprehension and analysis of human emotions. It presents a multi-granular method, which enables it to identify emotions at several levels, ranging from basic ones (such as joy, rage, or fear) to more complicated or sophisticated emotional states (such as nostalgia or jealousy). The approach emphasizes on being robust across several datasets, context-adaptable, and explainable—that is, enabling users to comprehend why the AI came to a particular emotional conclusion. To better capture emotional nuances, Emotion AWARE combines machine learning, linguistic features, and contextual clues. Additionally, it addresses one of the main issues in emotional computing—the "black box" nature of AI models—by emphasizing interpretability and openness. In order to better capture emotional nuances, it also incorporates machine learning, linguistic elements, and contextual signals. Additionally, it addresses one of the main issues in emotional computing—the "black box" nature of AI models—by emphasizing interpretability and openness. The study shows that multi-level emotion representation improves user trust and accuracy in emotion-aware technologies. Essentially, Emotion AWARE opens the door to more human-like emotional intelligence in AI systems by offering a versatile, understandable framework that links emotion recognition, context comprehension, and interpretability (Gamage et al., 2024).

Paper 5: Eye Gaze and Self-Attention: How Humans and Transformers Attend Words in Sentences

The paper's authors contrast the self-attention distributions of transformer language models (like BERT) with human attention patterns as determined by eye tracking while reading. They looked into whether the words that people focus on (dwell time, number of fixations) matched the terms that models give high attention weights. They demonstrate moderate to substantial correlations between human gaze and early-layer transformer attention using a range of eye-tracking datasets and pretrained transformers. This implies that despite having extremely distinct internal mechanics, transformer models may, in certain respects, resemble human reading attention patterns.

Additionally, they discover that the correlation's strength is not just influenced by the model's size or number of parameters, suggesting that architectural elements or training data may be more significant. (Bensemann et al., 2022)

Paper 6: Emotion-Aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine grained emotion recognition

Two key innovations are added to the transformer model RoBERTa for textual emotion recognition: a TF-IDF based gating mechanism that suppresses low-information tokens prior to encoding, and an Emotion-Specific Attention (ESA) layer that re-weights attention on emotionally salient tokens. While generalizing to external datasets (88.03% and 65.67% on large-scale and noisy corpora, respectively), the model obtained 96.77% accuracy and a weighted F1-score of 0.97 using a primary dataset of six emotions, an improvement of almost 10 percentage points over baseline RoBERTa. The ESA and TF-IDF components together enhanced focus on important emotional cues and decreased noise, as demonstrated by visualizations and ablation studies. The methodology issues with figurative language (sarcasm, irony), handles emotions as single labels rather than multi-label, and hasn't been tested on low-resource languages or multimodal data, according to the authors. (Alqarni et al., 2025)

Critical Analysis

When taken as a whole, the papers provide a strong basis for comprehending how contemporary transformer-based models interpret linguistic emotional and cognitive data. Devgan (2023) and Alqarni et al. (2025) show that contextualized transformers outperform baseline models in capturing fine-grained emotional nuances, particularly when reinforced by emotion-specific attention or TF-IDF gating. This is further demonstrated by Guo and Choi (2021) and Zhang et al. (2022), who show that emotional structure can be directly encoded into learnt embeddings, demonstrating that affective meaning arises organically inside neural representations rather than only as a result of external annotation. Gamage et al. (2024) offer a more comprehensive framework called Emotion AWARE, which emphasizes the significance of multi-granular and interpretable emotion modeling by combining adaptability, robustness, and explainability. In the meantime, Bensemann et al. (2022) present evidence of cognitive alignment, demonstrating a correlation between human eye-gaze patterns and transformer self-attention, indicating some similarities between human and artificial attentional processes. When evaluated as entirety, this research demonstrate that language models are capable of identifying and representing affective information, that attention may resemble certain features of human cognition, and that architectural improvements can increase interpretability and accuracy.

Connections

1. **Affect representation foundations:** They demonstrate how embeddings can map to psychological models (Plutchik, PAD) and encode emotional qualities.
2. **Modeling / mechanism references:** When extracting or perturbing the model to investigate the interaction between cognition and emotion, they offer techniques for contextual fine-tuning and attention gating.
3. **Explainability & multi-granularity:** serve as inspiration for examining emotions at various levels (basic to sophisticated) and providing evidence for the need to investigate the model at various granularities.
4. **Human-model cognitive alignment precedent:** In order to compare model signals to human cognitive models or proxies (attention, valence/arousal mapping), they empirically connect transformer attention to human attention (eye gazing).

Conflicts between the findings

These studies show significant constraints and conflicts despite their overall advancements. The majority of studies on emotion recognition ignore how thinking and attention affect emotion understanding, treating emotion as a categorization problem rather than a continuous, cognitively affected process. Embedding analyses, on the other hand, investigate affect representation but leave out the cognitive component, making it difficult to understand how emotional meaning interacts with attention or contextual cognition. Some bridge the gap between human and transformer attention, but they don't discuss how emotional valence influences this alignment. These gaps show that transformers lack a cohesive framework that links emotion and cognition. Additionally, inconsistent datasets, language complexity, and model scale introduce unpredictability in results, highlighting the necessity of cross-model evaluation with a psychological foundation.

Synthesis matrix for Literature Review

Study	Authors	Year	Research Design	Key findings/Results	Methodology	Limitations
Contextual emotion recognition using transformer-based models	Devgan, A., & Singh, S.	2022	Experimental	The evaluation metrics included F1 score, recall, accuracy, and precision. The model was trained for five epochs on the dataset, and the testing accuracy was 0.9310. As the number of epochs increased, the training accuracy also improved while the training loss reduced. Examined into by optimizing BERT on the dataset, Contextual Emotion Recognition achieved good accuracy and F1-score. This strategy's efficacy highlights the need of contextual knowledge for emotion recognition tests.	The dataset was compact, with 2000 rows in the test CSV file and 16,000 rows in the training dataset. According to the statement, the dataset contains six different emotions: fear, love, rage, surprise, sadness, and joy. Dataset collection, preprocessing, tokenization, fine-tuning, evaluation, and inference comprised the pipeline. Even though the pipeline was mentioned, the study did not go into great depth about each step.	Although the study successfully illustrates how transformer-based models might enhance contextual emotion recognition, it is still narrowly focused and seemingly psychological; it doesn't examine the reasons behind or methods by which language models encode emotions or the relationship between those emotional cues and human cognition.

Representing affect information in word embeddings	Zhang, Y., Chen, W., Zhang, R., & Zhang, X.	2022	Experimental	The affect-enhanced embeddings demonstrated better emotion sensitivity and stronger alignment with human emotion ratings, outperforming baseline models such as Word2Vec, GloVe, and BERT when tested on emotion detection, sentiment analysis, and affect prediction tasks.	The research combines normal text corpora with affective lexicons, which score words according to valence, arousal, and dominance, to produce embeddings that capture both semantic and emotional links between words. Distributional semantics and affective scores are used in the model's training process. Affective constraints are used to fine-tune word vectors based on emotional associations after the model has been pre-trained on a huge amount of unlabeled text.	Even though the study effectively adds emotional meaning to word embeddings, it is still emotion-only, static, and linguistically superficial, lacking contextual flexibility and cognitive insight.
Enhancing cognitive models of emotions with representation learning	Guo, Y., & Choi, J. D.	2021	Experimental	On the 32-emotion classification task in the Empathetic Dialogue dataset, the model achieves state-of-the-art accuracy.	The study applies a multi-head probing model on top of a contextualized embedding encoder to process text labeled with fine-grained emotion	Only the depiction of emotions was the emphasis; cognitive functions like reasoning, memory, and attention were not. They

				An emotion graph reflecting hierarchical links among emotions (e.g., related emotions cluster) can be derived thanks to the layer-wise probing, which demonstrates that specific layers collect emotion-specific information. The embeddings show some agreement between learnt emotion representations in text and human-theoretical emotion structure, supporting the creation of an emotion wheel comparable to human psychology models (e.g., Plutchik's). The potential of the embedding space for more extensive affective representation	categories, allowing the learnt representations to be interpreted. The empathic dialogue dataset is then used to train the algorithm, and its classification accuracy is assessed. The representation structure is then examined using layer analysis, which produces a "emotion wheel" mapping to human emotion theory and a "emotion graph" that illustrates the relationships between emotion classes. Additionally, it is demonstrated that these embeddings can be applied to different emotion models to supplement missing values.	employ contextual embeddings at the sentence level, but rather than looking at the relationship between embeddings and cognitive-like operations, the research focuses on the categorization and clustering of emotions. There are just 32 emotions on a dialogue dataset, making the dataset and job (emotion categorization) somewhat limited. Cognitive-emotional interaction may be more complicated in the real world. They make an effort to be interpretable (using emotion graphs and wheels), however they don't thoroughly examine cognitive layer dynamics or attention
--	--	--	--	--	--	---

				tasks is demonstrated by its application in emotion augmentation, which fills in gaps in emotion models (such as missing values in PAD).		mechanisms. Although the work does not connect to neuroscientific evidence or human cognition models (e.g., how emotion effects cognition or how emotion–cognition interactions manifest in humans), the agreement with human emotion theory is encouraging.
Emotion Aware: an artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis	Gamage, G., De Silva, D., Mills, N., Alahakoon, D., & Manic, M.	2024	Experimental	Across several datasets and all evaluated granularity levels (2, 8, and 14 emotions), the Emotion AWARE framework consistently beat state-of-the-art techniques in emotion recognition. Compared to solely LLM-based or purely lexicon-based approaches, the hybrid technique of combining	The authors created a framework consisting of three modules: A pretrained language model (such as BERT) is used in Module 1 (Emotion Language Model Finetuning), which refines it for emotion detection; Module 3 (AWARE Core) combines the refined model and lexicon outputs to create emotion assemblies at	Metaphorical or figurative language that subtly conveys emotion, as well as highly technical, jargon-rich, or informal emoji-based phrases, continue to be challenging for the framework. Full out-of-the-box generalizability is limited by the framework's requirement for domain-specific term lists for lexicon generation or

				lexicon heuristics with refined LLMs offered superior resistance to modifiers like negations and intensifiers and higher flexibility to domain-specific vocabularies. The framework improves the transparency of emotion detection judgments by listing constituent terms and intensity scores for each emotion. The multi-granular design showed that intensities, as opposed to single labels, give richer emotional profiles and that higher granularity (14 emotions) is achievable.	various granularity levels (2, 8, and 14 emotions) with intensity scores. Module 2 (Emotion Lexicon Generation) creates or modifies lexicons of emotion terms, including modifiers (negations, intensifiers, and inhibitors). The method makes the system both knowledge-driven and data-driven by combining deep learning (LLM fine-tuning) with lexicon-based heuristics. The framework's flexibility, robustness, and explainability are then tested in nine empirical investigations using eight different datasets spanning a range of areas. For classification tasks, metrics such as accuracy, precision, recall, and F1-score were employed.	additional fine-tuning for highly specialized domains, even when domain adaptability is handled. The intensity score and constituent phrase extraction still rely on heuristics, which may not adequately reflect the cognitive or contextual reasoning behind emotional expressions (e.g., internal reasoning, underlying motives), even though explainability has increased.
Eye-gaze and self -attention:	Bensemman, J., Peng, A.	2022	Experimental	The study discovered a	The study makes use of many eye-	Although it offers insightful

how humans and transformers attend words in sentences	Y., Benavides-Prado, D., Chen, Y., Tan, N., Corballis, P. M., Riddle, P., & Witbrock, M.			strong correlation between human eye look during reading and the attention patterns of transformer models, especially in the early levels. Larger models, however, do not always correspond better with human attention, suggesting that model size is not the primary determinant. The dataset also affects the correlation's strength; simpler texts exhibit greater alignment. A similar feeling of sentence importance is shown by the fact that both humans and transformers concentrate more on content words (such as nouns) than on function words.	tracking datasets and pre-trained transformer models, including BERT-Base, BERT-Large, and RoBERTa. To create an attention score for every word, self-attention weights from every token in every layer have been retrieved and averaged. Each token's attention score has been mapped to the human gaze measure of the matching word. They calculate spearman correlation coefficients between model attention distributions across layers and models and human attention signals. The correlation strengths are evaluated between datasets, model sizes and families, and layers. For qualitative understanding, attention	information about attention alignment, its focus is restricted to perceptual attention and ignores the more profound emotional or cognitive mechanisms underlying language comprehension. The analysis ignores the ways in which emotions affect attention or the ways in which cognitive processes like reasoning and context integration aid in the creation of meaning.
---	--	--	--	---	--	--

					visualizations are compared to gaze distributions.	
Emotion Aware RoBERTa enhanced with emotion specific attention with TF-IDF gating for fine grained emotion recognition	Alqarni, F., Sagheer, A., Alabbad, A., & Hamdoun, H.	2025	Experimental	On the primary dataset, the improved model outperformed baseline RoBERTa and other transformers with an accuracy of 96.77% and a weighted F1-score of 0.97 (improvement around +9.7% to +10.9%). The model demonstrated good but falling generalization in noisy environments, achieving 88.03% accuracy on a large-scale corpus and 65.67% on a smaller noisy social media dataset. Attention visualizations reveal that the ESA layer concentrates more on emotionally salient tokens than baseline	A pretrained RoBERTa backbone for textual emotion recognition is where the authors start. Two significant improvements are added: a TF-IDF based gating mechanism that filters or down-weights low-information tokens before feeding them into RoBERTa, and an Emotion-Specific Attention (ESA) layer that re-weights attention to emotionally salient tokens. Text normalization (slang, informal text), tokenization using the RoBERTa tokenizer, oversampling of underrepresented emotion classes, and dynamic batching are examples of preprocessing. The model is evaluated using	The study restricts the capture of overlapping or mixed feelings (e.g., melancholy + anxiety) by treating emotion identification as single-label categorization. Even with ESA improvements, the model still has trouble grasping complex linguistic phenomena like irony, sarcasm, and indirectly communicated emotion. The model's performance in low-resource languages or code-mixed text has not been tested, despite the usage of several datasets. Despite the advancements, the architecture is still computationally

				attention; ablation results demonstrate that both ESA and TF-IDF gating contribute to performance gains. Computational optimizations (FP16 + Automatic Mixed Precision) have been observed to reduce inference time and memory usage, indicating deployment readiness.	two external datasets (one large-scale, one noisy social media) and a primary dataset (6 emotion classes) to demonstrate generalization. Accuracy, weighted F1-score, Matthews Correlation Coefficient (MCC), and ablation studies (with and without ESA/TF-IDF) are among the metrics.	demanding when compared to lighter models, which could restrict its use in real-world scenarios with limited resources. Rather than delving deeper into emotion-cognition relationships, the work focuses on emotion identification.
--	--	--	--	---	--	---

References

1. Devgan, A. (2023). Contextual emotion recognition using transformer-based models. TechRxiv. <https://doi.org/10.36227/techrxiv.23804025.v1>
2. Zhang, Y., Chen, W., Zhang, R., & Zhang, X. (2022). Representing affect information in word embeddings. *Experiments in Linguistic Meaning*, 2, 310-321. <http://dx.doi.org/10.3765/elm.2.5391>
3. Guo, Y., & Choi, J. D. (2021). Enhancing cognitive models of emotions with representation learning. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2021)* (pp. 141-148). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.cmcl-1.18>
4. Gamage, G., De Silva, D., Mills, N., Alahakoon, D., & Manic, M. (2024). Emotion AWARE: an artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis. *Journal of Big Data*, 11(1), Article 93. <https://doi.org/10.1186/s40537-024-00953-2>
5. Alqarni, F., Sagheer, A., Alabbad, A., & Hamdoun, H. (2025). Emotion-Aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-99515-6>
6. Bensemann, J., Peng, A. Y., Benavides-Prado, D., Chen, Y., Tan, N., Corballis, P. M., Riddle, P., & Witbrock, M. (2022). Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2022)* (pp. 75-87). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.cmcl-1.9>