Integrating Plutchik's Theory with Mixture of Experts for Enhancing Emotion Classification

Dongjun Lim
Sungkyunkwan University
Suwon, South Korea
flame1220@skku.edu

Yun-Gyung Cheong*
Sungkyunkwan University
Suwon, South Korea
aimecca@skku.edu

Abstract

Emotion significantly influences human behavior and decision-making processes. We propose a labeling methodology grounded in Plutchik's Wheel of Emotions theory for emotion classification. Furthermore, we employ a Mixture of Experts (MoE) architecture to evaluate the efficacy of this labeling approach, by identifying the specific emotions that each expert learns to classify. Experimental results reveal that our methodology improves the performance of emotion classification.

1 Introduction

Emotion is essential in human life, having influence on our thoughts, behaviors, and communication. Recognizing the paramount importance of emotions, researchers have made significant efforts to analyze and understand them (Picard, 1997). A particularly important area of this research is emotion recognition in text, as it forms a substantial part of our daily interactions, including email and Social Network Service (SNS).

While sentiment analysis, categorizing text as positive, negative, or neutral, has advanced significantly, recognizing the full spectrum of emotions in text–such as *joy*, *anger*, *sadness*, and *fear*–remains a challenging task. Mao et al. (2023) report that RoBERTa large with HG-F24 achieved 84.7% accuracy on sentiment analysis of Amazon product reviews but only 40.9% accuracy in emotion detection using a Twitter (X) dataset.

Previous research utilizing deep learning technology has demonstrated significant promise in extracting emotions from text (Yu et al., 2018; Baziotis et al., 2018; Ying et al., 2019; Li and Xiao, 2023; Alhuzali and Ananiadou, 2021). Recently, Chen et al. (2023) conducted a study analyzing the role of emotions in controversial Reddit comments using language models. He et al. (2024) systematically measured the affective alignment of language models (LMs) by comparing LM-generated responses to SNSs on two socio-political issues. However, these studies face challenges like sampling bias and subjective annotation. For instance, Chai et al. (2024) note that existing multi-label text classification models lack the ability to generalize complex concepts. Ahanin et al. (2023) argue that current methods overlook the sentiment polarity of words.

To tackle the problems in emotion annotation, we introduce a new labeling approach. Our primary objective is to enhance the expressiveness of emotion labels by applying Plutchik's Wheel of Emotions and Diagram of Emotion Dyads. Furthermore, we employ a Mixture of Experts (MoE) framework for emotion classification, which identifies the specific emotion that each expert in the model is best at classifying. This approach seeks to validate the improved classification performance and specialization of experts in distinct emotional categories.

The key contributions of this research are listed as follows:

- We propose a new emotion labeling method based on Plutchik's wheel of emotions theory.
- We leverage MoE that is trained on basic emotions and learns to classify complex emotions effectively.
- We conducted experiments to show the efficacy of the proposed method. The results demonstrate that our approach can effectively improve the performance of emotion classification tasks, especially for emotions that are typically harder to classify with traditional methods.

The structure of the paper is organized as follows. Section 2 provides a review of related work. Section 3 outlines our approach. Section 4 details

*Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 857–867
November 12-16, 2024

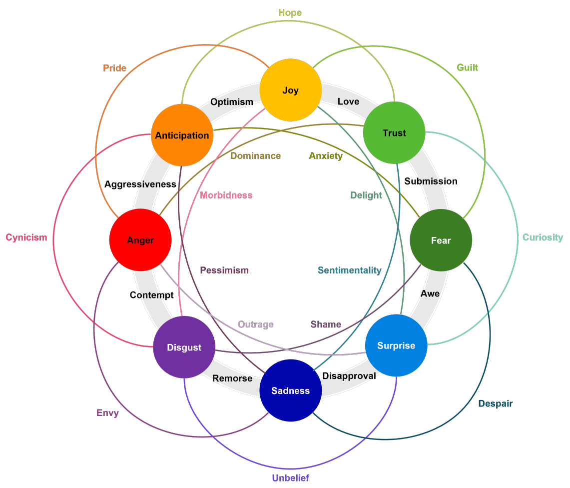


Figure 1: Plutchik's Diagram of Emotion Dyads. Depicting the primary, secondary, and tertiary dyads formed by mixing the eight basic emotions (Plutchik, 1991, 2000).

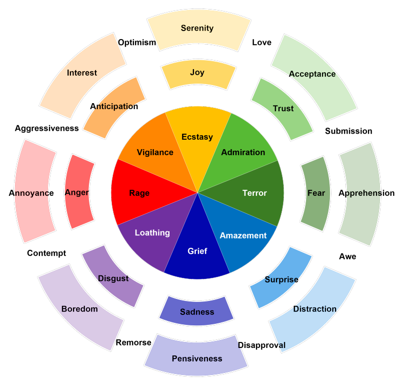


Figure 2: Plutchik's Wheel of Emotions. The eight emotions are represented within the color spectrum, showing their mild and intense variations (Plutchik, 1988).

the experimental design. Section 5 discusses the results, and Section 6 provides an in-depth analysis. The final section concludes with future research.

2 Related Work

2.1 Affective Computing

Emotions are physical and mental states induced by neurophysiological changes, often associated with specific thoughts, feelings, behavioral responses, and varying degrees of pleasure or displeasure (Damasio, 1998; Ekman and Davidson, 1994; Panksepp, 2004). They intertwine with mood, temperament, personality, disposition, and creativity (Averill, 1999). Recent research across psychology, medicine, history, sociology, and computer science highlights the complexity and importance of understanding emotions.

Despite extensive research, there is no universally accepted definition of emotion (Cabanac, 2002; Clore and Ortony, 2008). Emotions are categorized into various affects corresponding to specific situations (Barrett, 2006), and numerous theories have been proposed, each offering distinct perspectives on emotional experiences (James, 1884; Candland, 2003).

Ekman has significantly advanced our understanding of basic emotions through his research on facial expressions (Ekman, 1984). He identified six fundamental emotions: *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise* (Ekman, 1992a,b; Miller, 2016). Later, he expanded this list to include *amusement*, *contempt*, *contentment*, *embarrassment*, *excitement*, *guilt*, *pride in achievement*, *relief*, *satisfaction*, *sensory pleasure*, and *shame*, recognizing emotions not expressed solely through facial muscles (Ekman, 1999).

Our labeling method relies on Plutchik's emotion theories (Plutchik, 2000, 1988), which define eight basic emotions, grouped as *joy* versus *sadness*; *anger* versus *fear*; *trust* versus *disgust*; and *surprise* versus *anticipation*. These basic emotions can combine to form complex emotions, as depicted in Figure 1; for instance, the complex emotion *love* is formed by *joy* and *trust*, while *remorse* is a mix of *disgust* and *sadness*. These complex emotions may arise from cultural conditioning or associations combined with the basic emotions. He further introduced twenty-four 'Primary,' 'Secondary,' and 'Tertiary' dyads, representing different emotion combinations, and noted that emotions can vary in intensity from mild to intense (Plutchik, 1991; Turner, 2000). As illustrated in Figure 2, for instance, *annoyance*, *anger*, and *rage* fall within the same category with different intensities.

2.2 Mixture of Expert

The Mixture of Experts (MoE) method divides complex problems into multiple sub-problems, using specialized models (i.e., experts) to address each sub-problem. MoE utilizes a gating network to combine the outputs of each expert model, selecting the most suitable expert for a given input. This approach is particularly useful for datasets with diverse characteristics, enhancing model performance and computational efficiency.

Eigen et al. (2013) introduced the idea of using multiple MoEs, each with its own gating network, as part of a deep model. This approach is more powerful since complex problems may contain many sub-problems, each requiring different experts. They also suggest that introducing sparsity could transform MoE into a tool for computational efficiency. Shazeer et al. (2017) proposed a new type of general-purpose neural network component: a Sparsely-Gated Mixture-of-Experts Layer (MoE). This method uses Noisy top-k gating, which adds sparsity and noise to the Softmax Gate used in the MoE architecture (Jordan and Jacobs, 1994), selecting the top k values among the experts to produce the output. There are numerous other attempts to improve the gate network (Clark et al., 2022; Hazimeh et al., 2021; Zhou et al., 2022).

Lepikhin et al. (2020) replaced the Transformer Encoder's FFN layer with MoE, distributing experts across devices. This had the drawback of slower speeds when computations concentrated on a single expert. Fedus et al. (2022) improved this by limiting each token to one expert (k=1) and restricting the number of tokens per expert. Jiang et al. (2024) used an MoE structure with Top-k Gating and SwiGLU as experts within the Mistral model's Transformer block, improving performance across tasks and showing each expert specialized in specific tasks.

3 Method

This section describes our proposed method for emotion classification, utilizing the new labeling method based on Plutchik's emotion theory and the implementation of the MoE structure in our model.

3.1 Plutchik Labeling

We redefine the emotion labels of any dataset we wish to use, based on the work of Plutchik (2000, 1988). Data labeled with our method are termed "Plutchik Labeling" and and those without it as "Normal Labeling." The Plutchik Labeling process follows the following rules:

- Labels corresponding to the eight basic emotions in Plutchik's emotion theory are retained.
- Labels corresponding to primary, secondary, and tertiary dyads of the eight basic emotions are decomposed into their constituent emotions before labeling.
- Emotions that are combinations of opposite emotions are similarly decomposed into their constituent emotions before labeling.
- Mild and intense emotion labels are relabeled as the corresponding basic emotions.

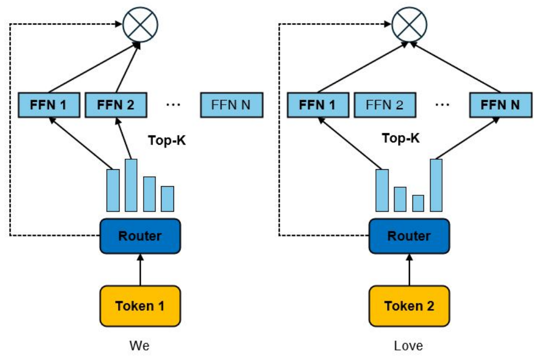


Figure 3: The Structure of Top-k MoE FFN.

While Plutchik's emotion theory also hints at the existence of triads (Plutchik, 1991), these dataset did not provide sufficient detail on these emotions. Therefore, our study does not consider the triads, higher-order combinations, or the intensity of emotions.

3.2 Mixture of Emotion Expert

We aim to apply Mixture of Experts (MoE) to each model to determine whether each expert can be trained as a specialist in individual emotions. As previously mentioned, there are several gating methods that connect inputs to specific experts. Following the approach in Jiang et al. (2024), we selected the k most relevant experts for each token. The reason for experimenting with multiple values of k instead of fixing it is to account for complex emotions such as love and optimism, which are described as mixtures of several basic emotions according to Plutchik (2000, 1988). This consideration is crucial when tokens contain complex emotions.

For the implementation of MoE, we referred to Mixtral (Jiang et al., 2024). The MoE structure used in Mixtral determines the output for a given input x by taking a weighted sum of the expert networks' outputs, with weights provided by the gating network. This is efficiently implemented using a softmax over the Top-K logits of a linear layer. A brief overview of the MoE Layer is provided in Figure 3.

To compare how well the model understands emotions when MoE is applied, we used the FFN network of the base model as experts. To observe the performance changes with minimal parameter modifications, we replaced the FFN network in the last transformer block of each model with an MoE structure.

Original Emot.	Relabeled Emot.
Love	Joy, Trust
Optimism	Anticipation, Joy
Pessimism	Anticipation, Sadness

Table 1: Rules for relabeling compound emotions as the corresponding basic emotions in SemEval-2018.

Original Emot.	Relabeled Emot.
Admiration	Trust
Annoyance	Anger
Confusion	Anticipation, Surprise
Curiosity	Surprise, Trust
Disappointment	Sadness, Surprise
Disapproval	Sadness, Surprise
Excitement	Fear, Joy
Grief	Sadness
Love	Joy, Trust
Optimism	Anticipation, Joy
Pride	Anger, Joy
Remorse	Disgust, Sadness

Table 2: Rules for relabeling compound, mild, and intense emotions as the corresponding basic emotions in GoEmotions.

4 Experiments

This section details the experimental design for evaluating the effectiveness of the proposed method in multi-label emotion classification.

4.1 Experimental Setup

Our experiments utilize two transformer-based models, Llama-2(Touvron et al., 2023) and Mistral(Jiang et al., 2023), each with 7 billion parameters, chosen for their effectiveness across various domains (Hou et al., 2024; Yu et al., 2024; Gruver et al., 2023). The unmodified versions of these models served as baselines for comparison. We accessed these models via the Hugging Face API and fine-tuned them using Q-LoRA(Dettmers et al., 2024). For all experiments, we used the same hyperparameters except for the k value. Performance was evaluated by averaging the results over five runs for each setting. Detailed hyperparameter configurations are provided in Section A.1.

4.2 Labeling for Building Datasets

We chose the evaluation datasets based on the following criteria: (1) inclusion of all 8 basic emotions from Plutchik's wheel, or (2) inclusion of emotions corresponding to Plutchik's 'Primary',

Emotion	train	valid	test
Anger	2544	315	1101
Anticipation	978	124	425
Disgust	2602	319	1099
Fear	1242	121	485
Joy	2477	400	1442
Love	700	132	516
Optimism	1984	307	1143
Pessimism	795	100	375
Sadness	2008	265	960
Surprise	361	35	170
Trust	357	43	153

Table 3: Emotion distribution across train, validation, and test sets for SemEval-2018 with Normal labeling.

Emotion	train	valid	test
Anger	2544	315	1101
Anticipation	3216	453	1688
Disgust	2602	319	1099
Fear	1242	121	485
Joy	2991	454	1669
Sadness	2266	292	1049
Surprise	361	35	170
Trust	975	161	621

Table 4: Emotion distribution across train, validation, and test sets for SemEval-2018 with Plutchik labeling.

Emotion	train	valid	test
Admiration	4130	488	504
Anger	1567	195	198
Annoyance	2470	303	320
Confusion	1368	152	153
Curiosity	2191	248	284
Disappointment	1269	163	151
Disapproval	2022	292	267
Disgust	793	97	123
Excitement	853	96	103
Fear	596	90	78
Grief	77	13	6
Joy	1452	172	161
Love	2086	252	238
Optimism	1581	209	186
Pride	111	15	16
Remorse	545	68	56
Sadness	1326	143	156
Surprise	1060	129	141

Table 5: Emotion distribution across train, validation, and test sets for GoEmotions with Normal labeling.

Emotion	train	valid	test
Anger	3877	464	504
Anticipation	2944	360	336
Disgust	1334	164	179
Fear	1448	186	181
Joy	5801	707	669
Sadness	4928	643	607
Surprise	7472	944	951
Trust	8125	956	994

Table 6: Emotion distribution across train, validation, and test sets for GoEmotions with Plutchik labeling.

Top-k	SemEval-2018		GoEmotions	
	miF1	maF1	miF1	maF1
baseline	70.7	56.4	64.2	**58.7**
1	70.6	56.4	63.5	58.5
2	70.8	**57.0**	63.8	58.0
3	70.7	56.1	63.8	58.0
4	**70.8**	55.9	**64.3**	58.7
baseline	70.3	55.4	63.7	58.2
1	70.5	55.4	63.8	58.9
2	70.3	**55.5**	64.1	58.9
3	69.6	54.7	64.0	59.2
4	**70.7**	54.6	**64.2**	**59.3**

Table 7: F1 scores of the models with Normal Labeling. Upper: Llama2, Lower: Mistral

'Secondary', and 'Tertiary' dyads, which, when decomposed, satisfy criterion 1. As a result, we selected SemEval-2018 (Mohammad et al., 2018) and GoEmotions (Demszky et al., 2020).

SemEval-2018 contains tweets, each labeled with one or more of 11 emotions, or marked as Neutral. GoEmotions consists of 58K Reddit comments from 2005 to 2019, each labeled with one or more of 27 emotions, or marked as Neutral. The rules for applying Plutchik labeling to these datasets are detailed in Tables 1 and 2.

For a fair comparison, we excluded data for emotions not covered by Plutchik's 8 basic emotions or their dyads, as well as Neutral, in all experiments. The final datasets are detailed in Tables 3, 4, 5, and 6. We fine-tuned the classification models using the training sets and evaluated their performance on the test sets.

5 Results

5.1 Main Results

Tables 7 and 8 present the F1-scores of our proposed methods on two datasets. Table 7 shows the performance for different k values when applying MoE in Normal Labeling. For SemEval-2018, the macro-F1 indicates the model exceeds baseline performance at k=2, achieving the highest performance. In GoEmotions, the Mistral model surpasses the baseline across all k values, peaking at k=4, while the Llama2 model underperforms at all k values. The micro-F1 shows the highest performance at k=4 in all cases.

Overall, SemEval-2018 shows a consistent trend in macro-F1 changes with varying k values, unlike GoEmotions. This inconsistency, as shown in Table 5, is due to significant label imbalance in GoEmotions. Elbayad et al. (2023) and Fedus et al. (2022) explain that MoE models tend to overfit on low-resource data, suggesting that the experts in the MoE model failed to learn effectively for certain emotions due to extreme imbalance. Additionally, *grief* and *pride* have significantly fewer test samples, leading to high variance in performance metrics. Thus, performance comparisons using macro-F1 in GoEmotions may not be accurate.

Table 8 presents the performance of MoE with Plutchik Labeling varying the k values . With SemEval-2018, the highest macro-F1 was obtained at k=3, outperforming the baseline model. In GoEmotions, the Mistral model achieved the highest score at k=4, while the Llama2 model exceeded the baseline at k=1. The highest micro-F1 score was generally obtained at k=3, except for the Mistral model on GoEmotions, which showed different patterns.

Plutchik Labeling resulted in more stable and superior performance than Normal Labeling, especially in GoEmotions, mitigating the effects of severe label imbalance. The MoE-trained model consistently outperformed the baseline model across various k values.

Figure 4 depicts the changes in macro-F1 performance across both datasets with varying k values. When applying Plutchik Labeling, there is a significant improvement in performance compared to Normal Labeling, both in the baseline and all MoE configurations. Notably, in SemEval-2018, when k is set to 1, the performance improvement with Plutchik Labeling is less pronounced compared to the baseline and other k values. This suggests that selecting two or more experts in SemEval-2018 allows for better interpretation of emotions.

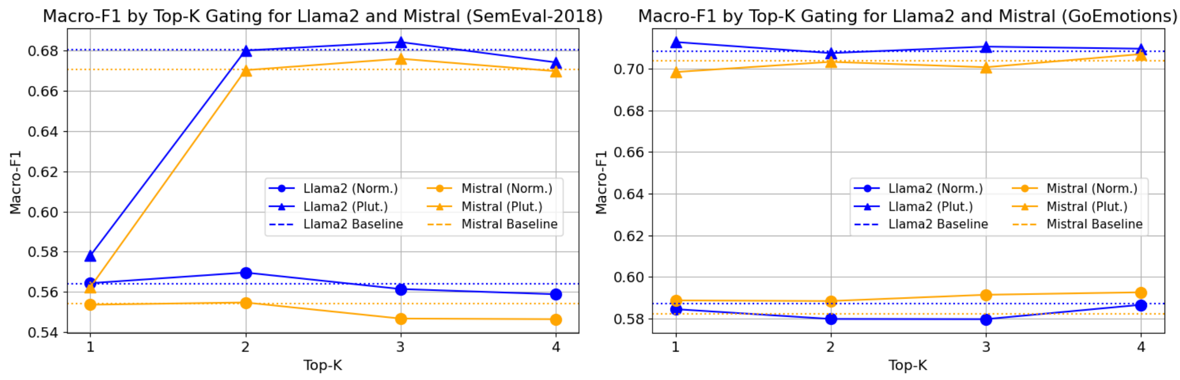


Figure 4: The macro-F1 scores of the MoE model across each datasets, k values, and labeling methods.

Top-k	SemEval-2018		GoEmotions	
	miF1	**maF1**	**miF1**	**maF1**
baseline	74.9	68.0	75.6	70.9
1	61.2	57.8	75.7	**71.3**
2	74.7	68.0	75.6	70.8
3	**75.0**	**68.4**	**75.8**	71.1
4	74.6	67.4	75.7	71.0
baseline	74.4	67.1	75.0	70.4
1	60.6	56.2	74.5	69.8
2	74.7	67.0	74.9	70.3
3	**74.9**	**67.6**	74.6	70.1
4	74.6	67.0	**75.1**	**70.7**

Table 8: F1 scores of the models with Plutchik Labeling. Top: Llama2, Bottom: Mistral.

Weak Emot.	**Llama2**		**Mistral**	
	Norm.	Plut.	Norm.	Plut.
ANT	24.0	66.8	24.3	69.4
PES	33.1	-	32.6	-
SUR	28.3	27.9	25.7	24.2
TRU	12.8	57.8	11.2	58.3
maF1	24.6	**42.7**	23.4	**50.6**

Table 9: F1-scores of underperforming emotions in SemEval-2018.

The optimal k values for classification varied across datasets, likely due to differences in the average number of labeled emotions. For instance, the SemEval-2018 dataset has 2-3 labels per instance, whereas the GoEmotions dataset has 1-2.

5.2 Underperforming Emotions

To assess the effectiveness of Plutchik Labeling, we tested whether it could enhance the classification of underperforming emotions, defined as those with F1-scores below 0.6 in the Normal Labeling dataset.

Table 9[1] presents the F1-scores for underperforming Emotions in SemEval-2018. When applying Plutchik Labeling, *pessimism* is decomposed into *anticipation* and *sadness*, resulting in the removal of the *pessimism* label. For basic Emotions, both *anticipation* and *trust* showed significant improvement in classification performance due to data augmentation. However, in the case of *surprise*, the transition from Normal Labeling to Plutchik Labeling did not benefit from data augmentation.

Table 10[1] presents the F1-scores for the underperforming emotions in GoEmotions. Basic emotions such as *anger*, *disgust*, and *surprise*—identified as underperforming emotions— demonstrated substantial improvement with Plutchik Labeling. However, many of the other underperforming emotions in GoEmotions are either complex emotions or represent varying intensities (mild or intense), making direct comparisons with Plutchik Labeling more difficult.

By comparing the macro-F1 scores of underperforming emotions between Normal Labeling and Plutchik Labeling in Tables 9 and 10, we observe a significant overall improvement in classification performance across both datasets. This enhancement suggests that our proposed method effectively improves the classification of emotions that are typically harder to classify accurately. We believe that this demonstrates the potential of Plutchik Labeling to enhance the robustness and accuracy of emotion classification systems.

[1] AN: Anger, ANO: Annoyance, ANT: Anticipation, CO: Confusion, CUR: Curiosity, DIS: Disappointment, DAP: Disapproval, DIG: Disgust, EXC: Excitement, GRF: Grief, LO: Love, OPT: Optimism, PES: Pessimism, PRI: Pride, REM: Remorse, SUR: Surprise, TRU: Trust

Weak Emot.	Llama2 Norm.	Llama2 Plut.	Mistral Norm.	Mistral Plut.
AN	57.0	66.4	51.2	65.0
ANO	45.3	-	45.2	-
CO	57.7	-	58.0	-
DIS	32.0	-	35.6	-
DAP	57.9	-	57.5	-
DIG	48.9	56.8	46.1	56.8
EXC	47.8	-	50.0	-
GRF	29.5	-	29.4	-
PRI	43.9	-	42.2	-
SUR	60.8	77.5	58.3	76.5
maF1	48.1	**66.9**	47.4	**66.1**

Table 10: F1-scores of underperforming emotions in GoEmotions.

Comp Emot.	llama2 baseline	llama2 k=2	mistral baseline	mistral k=2
LO	62.4	61.8	59.0	60.8
OPT	70.7	71.7	71.0	72.4
PES	33.1	37.7	32.6	37.3
maF1	55.4	**57.1**	54.2	**56.8**

Table 11: F1-scores of complex emotions in SemEval-2018.

Comp Emot.	llama2 baseline	llama2 k=4	mistral baseline	mistral k=4
CO	57.7	57.2	58.0	57.3
CUR	67.4	67.6	68.2	67.0
DIS	32.0	33.7	35.6	30.4
DAP	57.9	58.6	57.5	56.6
EXC	47.8	50.7	50.0	54.7
LO	83.3	83.9	84.2	85.6
OPT	68.7	70.3	69.8	69.9
PRI	43.9	38.2	42.2	41.9
REM	70.6	71.9	71.6	72.8
maf1	58.8	**59.1**	**59.7**	59.6

Table 12: F1-scores of complex emotions in GoEmotions.

5.3 Complex Emotions

To assess whether our MoE approach improves the classification of complex emotions, we compared the F1-scores of complex emotions between the baseline and MoE models under Normal Labeling. As similar trends were observed across various k values, we focused on the specific k values that showed the most significant improvement in macro-F1 scores for each dataset, relative to the baseline.

Table 11[1] presents the classification performance of complex emotions in SemEval-2018, comparing the baseline with the Top-2 MoE models. The MoE approach yielded a substantial improvement in macro-F1, significantly increasing the performance for *pessimism*, which was previously categorized as an underperforming emotion.

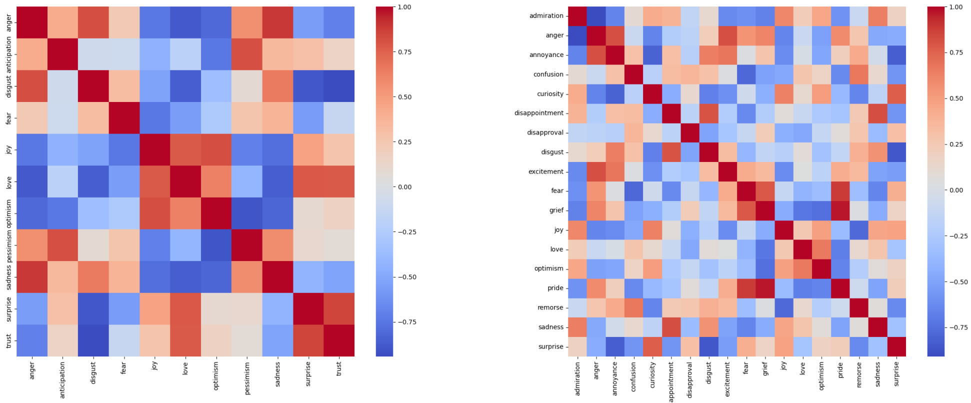
Table 12[1] presents the complex emotion classification performance of the baseline and Top-4 MoE models in GoEmotions. Based on macro-F1, Llama2 showed a slight improvement, while Mistral had a slight decline. Llama2's performance dropped for *confusion* and *pride*, whereas Mistral declined for *confusion*, *curiosity*, *disappointment*, *disapproval,* and *pride*.

Pride, with limited data samples, poses a challenge for performance improvement due to significant data imbalance. Other complex emotions, particularly those sharing elements with *surprise*, also face classification difficulties. According to Plutchik (1991), *confusion*, *curiosity*, *disappointment*, and *disapproval* overlap with *surprise*. However, Clore and Ortony (2013) argue that surprise is a cognitive state, not an emotion, as it lacks intrinsic valence and can manifest in both positive and negative contexts, depending on subsequent evaluations. This difference in perspective adds complexity to distinguishing *surprise* from related emotions that involve both cognitive and affective components. As a result, our study faced challenges applying the MoE model, which likely struggled to classify *surprise* and other complex emotions that range from neutral to evaluative.

6 Analysis

We investigated the relationships between emotions by analyzing the predominant expert selections for each. By tracking the output values of the Gate Layer in a Mixture of Experts (MoE) model, we identified which Experts were primarily selected for each emotion.

Our approach involved selecting Experts for each token and aggregating the selection proportions of the Top-k Experts per token for each input. The value of k corresponds to the Top-k used in the MoE, with the selection proportions for each token summing to 1. Inputs were grouped by their emotions labels, and the aggregate Expert selection proportions for each label were computed and standardized. Using these frequencies of Expert selections for each emotion, we plotted emotion-



(a) SemEval-2018

(b) GoEmotions

Figure 5: (a): Emotion correlations in Normal Labeling with Top-2 Gating. (b): Emotion correlations from in Normal Labeling with with Top-4 Gating.

emotion correlations to examine the relationships between emotions.

Figure 5a shows that *joy*, *love*, and *optimism* exhibit strong correlations, indicating that positive emotions are closely interconnected in the SemEval-2018 dataset. In contrast, *anger*, *sadness*, and *disgust* show strong positive correlations with each other, as well as with *fear* and *pessimism*, forming a cluster of negative emotions. Additionally, *optimism* and *pessimism*, as well as *love* and *sadness*, show strong negative correlations with each other, indicating that these emotions have opposite characteristics. Furthermore, *love* tends to have high correlations with *joy* and *trust*, *optimism* with *joy*, and *pessimism* with *anticipation* and *sadness*. These patterns also allow us to understand the similarities between complex emotions and their component basic emotions.

In GoEmotions, as shown in Figure 5b, *joy*, *love*, *optimism*, and *admiration* exhibit strong positive correlations, indicating their close interrelation as positive emotions. Conversely, *anger*, *annoyance*, *excitement*, *fear*, *grief*, and *pride* form a group of negative emotions, with *admiration* and *anger* showing a strong negative correlation, highlighting their opposing nature. Additionally, the complex emotions *disappointment* and *curiosity* correlate highly with *sadness* and *surprise*, respectively, while *anger* correlates strongly with *annoyance* and *sadness* with *grief*. These patterns reveal the similarities between complex emotions and their component emotions, as well as the relationships between basic emotions and their milder or more intense counterparts.

Overall, while the selection of Experts for each emotion does not perfectly align with Plutchik's emotion theory, the results show a significant degree of similarity. This suggests that our approach is effective for emotion analysis. These findings contribute to a deeper understanding of emotional interrelations and can aid in improving emotion prediction models.

7 Conclusion

Our approach, grounded in Plutchik's emotion theory and utilizing the MoE architecture, significantly enhances the performance of multi-label emotion classification tasks. The proposed methodologies were evaluated against baseline models, demonstrating significant improvements in classification accuracy. Notably, our approach excelled in identifying emotions that are traditionally difficult to classify and showed superior performance in recognizing complex emotions.

Moreover, the analysis of expert selection tendencies, based on emotion correlations, revealed that our model's behavior closely aligns with Plutchik's emotion theory. This alignment not only enhances classification accuracy but also provides a theoretically grounded insight into emotional interactions.

Thus, we believe that our research presents a robust framework for multi-label emotion classification, integrating psychological theories and advanced machine learning techniques in emotion recognition tasks. Future research could focus on refining the classification of mild and intense variations of emotions.

Limitations

This study acknowledges several limitations. First, utilizing Plutchik's emotion theory requires the dataset to include all eight basic emotions defined by the theory, posing a challenge for datasets lacking these emotions. Furthermore, excluding emotions not covered by Plutchik's emotion theory can be inefficient, making careful selection of datasets crucial. Future research could improve the labeling method by incorporating additional emotion models, such as the OCC model (Clore and Ortony, 2013).

Second, during the application of MoE, we encountered a known issue where tokens clustered around specific experts. This imbalance suggests the model may not fully leverage all experts. We plan to design a more sophisticated MoE structure to address this in the near future.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea grant funded by the Korean government(MSIT) (No. RS-2024-00357849), Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190421, AI Graduate School Support Program(Sungkyunkwan University)), the Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (No.RS-2024-00413839).

References

Zahra Ahanin, Maizatul Akmar Ismail, Narinderjit Singh Sawaran Singh, and Ammar AL-Ashmori. 2023. Hybrid feature extraction for multi-label emotion classification in english text messages. *Sustainability*, 15(16).

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

J R Averill. 1999. Individual differences in emotional creativity: structure and correlates. *J. Pers.*, 67(2):331–371.

Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46. PMID: 16430327.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Michel Cabanac. 2002. What is emotion? *Behavioural Processes*, 60(2):69–83.

D. Candland. 2003. *Emotion*. Core books in psychology. Authors Choice Press.

Yuyang Chai, Zhuang Li, Jiahui Liu, Lei Chen, Fei Li, Donghong Ji, and Chong Teng. 2024. Compositional generalization for multi-label text classification: A data-augmentation approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17727–17735.

Kai Chen, Zihao He, Rong-Ching Chang, Jonathan May, and Kristina Lerman. 2023. Anger breeds controversy: Analyzing controversy and emotions on reddit. In *Social, Cultural, and Behavioral Modeling*, pages 44–53, Cham. Springer Nature Switzerland.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. 2022. Unified scaling laws for routed language models. In *International conference on machine learning*, pages 4057–4086. PMLR.

Gerald Clore and Andrew Ortony. 2008. Handbook of emotions. *Appraisal theories: How cognition shapes affect into emotion*, pages 628–642.

Gerald L Clore and Andrew Ortony. 2013. Psychological construction in the OCC model of emotion. *Emot. Rev.*, 5(4):335–343.

Antonio R Damasio. 1998. Emotion in the perspective of an integrated nervous system1published on the world wide web on 27 january 1998.1. *Brain Research Reviews*, 26(2):83–86.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Paul Ekman. 1984. Expression and the nature of emotion.

Paul Ekman. 1992a. Are there basic emotions? *Psychological review*, 99(3):550–553.

Paul Ekman. 1992b. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Paul Ekman. 1999. *Basic Emotions*. John Wiley Sons, Ltd.

Paul Ekman and Richard J. Davidson, editors. 1994. *The Nature of Emotion: Fundamental Questions*. Oxford University Press USA.

Maha Elbayad, Anna Sun, and Shruti Bhosale. 2023. Fixing moe over-fitting on low-resource languages in multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14237–14253.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, pages 19622–19635. Curran Associates, Inc.

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. 2021. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems*, 34:29335–29347.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2024. Whose emotions and moral sentiments do language models reflect? *arXiv preprint arXiv:2402.11114*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

William James. 1884. II.—WHAT IS AN EMOTION ? *Mind*, os-IX(34):188–205.

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*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Michael I. Jordan and Robert A. Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6(2):181–214.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Jinfen Li and Lu Xiao. 2023. Multi-emotion recognition using multi-emobert and emotion analysis in fake news. page 128–135.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

Harold L. Miller. 2016. *The SAGE Encyclopedia of Theory in Psychology*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

J. Panksepp. 2004. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Series in Affective Science. Oxford University Press.

Rosalind W. Picard. 1997. Affective computing.

R. Plutchik. 1991. *The Emotions*. University Press of America.

Robert Plutchik. 1988. *The Nature of Emotions: Clinical Implications*, pages 1–20. Springer US, Boston, MA.

Robert Plutchik. 2000. Emotions in the practice of psychotherapy: Clinical implications of affect theories.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

J. Turner. 2000. *On the Origins of Human Emotions: A Sociological Inquiry into the Evolution of Human Affect*. Stanford University Press.

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 316–321.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, YU Jincheng, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Hyperparameter	**Value**
epoch	10
gradient_accumulation_steps	4
learning_rate	1e-4
warmup_ratio	0.1
max_grad_norm	0.3
weight_decay	0.001
batch_Size	8
quant_type	nf4
lora_r	8
lora_alpha	8
lora_dropout	0.1
num_expert	8

Table 13: Hyperparameter Settings for our experiments.

A Appendix

A.1 Hyperparameters

Table 13 shows the hyperparameter values applied to the models used in our experiments. Except for the k value, all hyperparameters were kept constant across all experiments. Each condition was tested five times.