Emotion Intensity Prediction using Transformer-Based Models

1st Damien Herlnata
School of Computer Science
Binus University
Jakarta, Indonesia
damien.herlnata@binus.ac.id

2nd Nabila Danisha Putri
School of Computer Science
Binus University
Jakarta, Indonesia
nabila.putri019@binus.ac.id

3rd Renji Earl Kurniawan
School of Computer Science
Binus University
Jakarta, Indonesia
renji.kurniawan@binus.ac.id

4th Bakti Amirul Jabar
School of Computer Science
Binus University
Jakarta, Indonesia
bakti.jabar@binus.ac.id

5th Jurike V. Moniaga
School of Computer Science
Binus University
Jakarta, Indonesia
jurike@binus.ac.id

***Abstract*—Emotion detection has been increasingly applied in various fields, including chatbots, virtual assistants and affective computing systems. However, most existing emotion detection models primarily focus on one task, which is to identify types of emotion expressed using emotion classification. While this is essential, a complete understanding of emotional expression also requires understanding the intensity of emotion expressed by the author. In this paper, we propose a Bidirectional Encoder Representations of Transformer (BERT) based model that can understand the type and strength of an emotion expressed through a single text. We approached this problem with two learning methods: Single-Task Learning, where each task is trained independently, and Multi-Task Learning, where both tasks are learned simultaneously through a shared encoder. Experiment results show that the multi-task learning model performs better than single-task learning**.

Keywords—Emotion classification, emotion intensity prediction, Single-Task Learning, Multi-Task Learning, BERT Transformer

I. Introduction

Emotion is a basic component in human communication; it influences how people communicate, make decisions and perceive the world [1]. While traditional emotion classification focuses on identifying text into different emotion categories—like the typical emotion categories proposed by Ekman and Plutchik [2] such as *joy, anger, sadness, fear, disgust, surprise*—it often overlooks the varying degrees to which these emotions are expressed. In various industries, such as mental health monitoring, virtual assistants, empathetic chatbot [3] and human-computer interaction, being able to understand just the type of emotion expressed is not enough. It's also crucial to understand how strongly those emotions are expressed.

Emotion intensity prediction aims to address this gap by measuring the magnitude or strength of an emotion expressed, providing more insight into emotional states beyond just emotion categories. This task is particularly challenging due to the subjective and context-dependent state of emotional expression [4], but recent advances in natural language processing and machine learning models have enabled promising approaches to tackle this problem.

Emotional expressions can be seen through various forms, including facial expressions [5]. Although the universality of emotional expression has been mentioned in several studies, language plays an important role in emotional design and how emotions are expressed [5]. Text not only communicates informative content, but also conveys information about the attitude and emotion expressed by the author [6]. Word choice can reflect emotional intensity—for example, both "angry" and "furious" expresses anger, but "furious" conveys a stronger emotion. Take the following two sentences:

(1) The boss is angry with the new intern

(2) The boss is very furious with the new intern

Both of these examples would be labelled with the emotion anger. However, it is clear that the second sentence expresses more magnitude/intensity of anger. This suggests that emotions are embedded in language, and by analyzing word choice, we can infer both the type and intensity of a person's feelings.

While most existing research focuses only on comparing different models for either emotion classification or intensity prediction separately, our research aims to extend Transformer-Based Models to jointly predict both the emotion category and the intensity level of the expressed emotion. Specifically, we compare the effectiveness of performing these tasks separately and jointly to examine whether combining both tasks in a single architecture leads to better overall performance.

The ability to accurately predict not just the emotion category but also the severity of those emotions can significantly enhance the emotional intelligence of AI systems [7] such as virtual assistant, mental health applications, sentiment analysis, et cetera. Through this research, we hope to contribute to the development of models that go beyond basic classification in order to bridge the gap between raw textual expression and nuanced emotional understanding, resulting in more natural and effective human-computer interaction.

II. Related Work

The ability to comprehend and analyze emotional expressions in textual data has emerged as a vital component in designing emotionally intelligent systems, particularly within domains such as affective computing [8][9], healthcare [10], customer service, and social media analytics. Traditionally, emotion classification has focused on categorizing text into a finite set of basic emotions [5]. However, contemporary research has moved beyond this categorical framework to address the prediction of emotion intensity—a more nuanced task that measures the magnitude of emotional expression. This shift is significant, as the strength of an expressed emotion often provides richer insight into user sentiment than the mere classification of emotional type [11][12].

To address this challenge, scholars have proposed a broad range of methodologies, spanning from rule-based lexical strategies and traditional machine learning algorithms to deep neural networks and transformer-based architectures. This chapter provides a comprehensive overview of key advancements and methodologies proposed in prior research on emotion intensity prediction, with a particular focus on textual data. These approaches collectively reflect the progression of the field from early statistical models to context-aware deep learning systems capable of capturing subtle emotional gradients embedded in natural language.

A landmark contribution in this area came from a study [13] that introduced the task of emotion intensity estimation in social media texts, specifically on Twitter. In this work, Saif M. Mohammad and Felipe Bravo-Marquez proposed an annotation technique known as Best-Worst Scaling (BWS), which enabled the construction of a fine-grained dataset focused on four primary emotions: anger, fear, joy, and sadness [13]. Annotators were instructed to assign continuous intensity scores (ranging from 0 to 1) to tweets based on the perceived strength of emotional expression. The study found that combining dense word embeddings with affective lexical resources yielded high performance in modeling emotional intensity, with top systems achieving strong correlation scores with human judgments.

Furthering this line of inquiry, another study [14] introduced a supervised regression framework that incorporated three core information sources: handcrafted affective lexicons, automatically derived lexical expansions, and a deep learning model employing a CNN-LSTM architecture. While each feature set independently yielded solid results, their combination produced a highly effective system for predicting emotion intensity on Twitter. Notably, the research demonstrated that customizing affective resources for the informal linguistic patterns typical of social media significantly enhanced prediction accuracy.

Building upon architectural innovations, a separate study [15] proposed a novel neural network design consisting of convolutional and fully connected layers arranged in parallel—a structure that was relatively unconventional for NLP tasks at the time. This model integrated transfer learning techniques with affective lexical features to improve intensity prediction. Results showed a substantial improvement over earlier state-of-the-art methods, highlighting the utility of parallel architectures in capturing diverse emotional signals.

Expanding on the multitask learning paradigm, another study [16] introduced a multi-task ensemble framework capable of handling both emotion classification and intensity prediction. The system integrated CNN, LSTM, and GRU networks alongside handcrafted features within a multilayer perceptron (MLP), resulting in a robust and generalizable architecture suitable for a wide array of emotion-related tasks. This approach underscored the synergistic potential of combining deep learning with engineered linguistic features.

The emergence of transformer-based models has further revolutionized the field, offering enhanced contextual understanding of emotional language. One study [1] leveraged BERT to detect contextual emotions, demonstrating that such models excel in identifying subtle emotional cues, especially when emotional states are implied rather than explicitly articulated. This advancement marked a key turning point in enabling models to engage in deeper emotional reasoning.

In alignment with this trend, another contribution [17] proposed DeepEmotex, a transfer learning-based model that fine-tuned BERT using emotion-labeled Twitter data. This approach showed strong performance across multiple domains and proved effective even with limited annotated datasets, reinforcing the utility of pre-trained transformers for low-resource emotion tasks.

Another research effort [18] introduced TA-MERT, a model that combined text augmentation techniques with transformer encoders to enhance emotion detection in conversational contexts. Evaluated on the MELD dataset, the model demonstrated improved robustness, particularly in handling class imbalance—an issue that commonly hinders performance in emotion classification tasks.

Recognizing the scarcity of annotated emotional datasets in languages other than English, one study [19] explored cross-lingual emotion intensity prediction. The researchers compared machine translation, multilingual embeddings, and unsupervised adaptation methods to transfer models trained in English to Spanish and Catalan. Interestingly, models with minimal reliance on parallel corpora often surpassed those that required extensive bilingual resources, highlighting the promise of lightweight cross-lingual techniques.

Furthering the field, another study [20] proposed a selective feature approach employing a Bi-LSTM model that emphasized emotionally salient words within a sentence. By integrating CNN, LSTM, and Bi-LSTM networks, the model generated vector representations focusing specifically on emotionally dominant tokens, which led to improved accuracy in classifying core emotions such as joy, fear, anger, and sadness. This method demonstrated the effectiveness of targeted attention toward emotionally charged words.

Lastly, a study [21] highlighted the importance of embedding contextual sensitivity into lexical features. Using a recurrent neural network (RNN), the researchers developed a mechanism that dynamically modulated word sentiment scores based on syntactic structures such as negation and intensifiers. Unlike traditional bag-of-words models, this context-aware approach enabled more nuanced sentiment interpretations and significantly improved classification performance.

III. Proposed Methodology

In order to build an emotion classification and intensity prediction model, we use a Transformer-based approach to model the type and strength of emotion expressed in a text, leveraging the capabilities of Bidirectional Encoder Representations from Transformers (BERT). BERT is a pre-trained language model that has shown a state-of-art performance on several NLP benchmarks showing its robustness across a wide range of tasks [22]. Due to its ability to capture rich contextual relationships between words through bidirectional context, BERT offers a solid and robust foundation for modeling both our task of emotion classification and intensity prediction.

Using BERT, we propose two learning approaches, namely, Single-Task Learning (STL) and Multi-Task Learning (MTL). In the STL setup, separate models are trained independently for the tasks of emotion classification and intensity prediction. Meanwhile, in the MTL setup, a single model is trained to solve both tasks simultaneously. Through these two approaches, we compare both results and aim to explore the potential benefits and trade-offs of MTL in improving generalization, using STL as a baseline for evaluation.

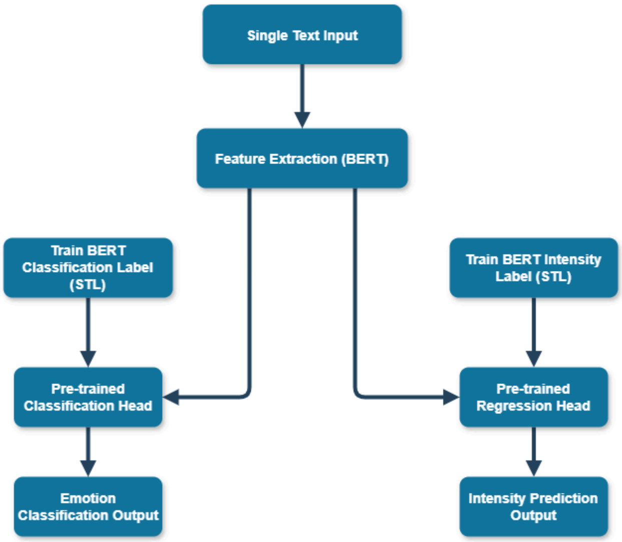

Figure 1: Proposed Single-Task BERT framework

A. *Single-Task Learning using BERT Transformer*

For the single-task learning approach, we employ the architecture illustrated in Figure 1. First, we develop two independent models: one for emotion classification and another for intensity prediction. Both models use bert-baseuncased as the base encoder, each fine-tuned using its respective set of labels without any parameter sharing between tasks. Each model also incorporates additional lexicon-based features concatenated with the BERT pooled output before passing through a dropout layer and final output head. The architecture is the same for both model pipelines, with only the final output head differing to suit their nature of the task.

For emotion classification, we use a linear classification head that outputs raw logits. These logits are then passed to a cross-entropy loss function, which internally applies softmax activation to produce a probability distribution between discrete labels over emotion categories. For intensity regression, we use a regression head that produces a single scalar output, scaled through tanh activation function to map the output to a normalized intensity score between 0 and 1.

The dataset is preprocessed to isolate the relevant task. For emotion classification, we only use text samples and emotion labels. For intensity prediction, we only use text samples and intensity scores. Each task is trained independently using the standard train-test-split. Cross-entropy loss is used to calculate the loss function for the classification task, and a differentiable Pearson correlation-based loss is used for the regression task.

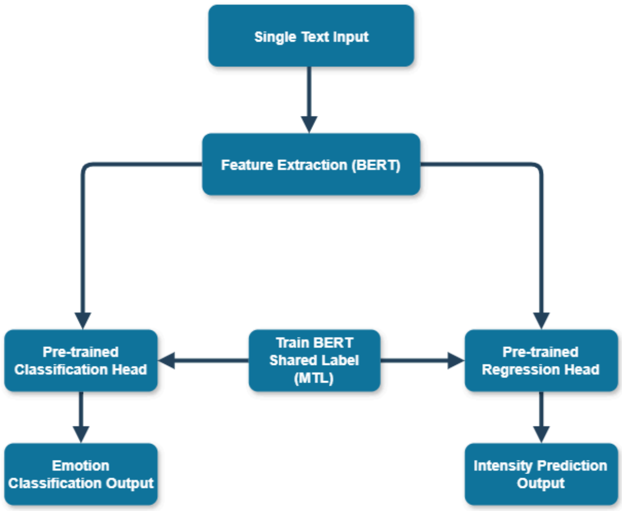

Figure 2: Proposed Multi-Task BERT framework

B. *Multi-Task Learning using BERT Transformer*

For the multi-task learning approach, we employ the architecture illustrated in Figure 2. First, we design a model that performs both emotion classification and intensity prediction simultaneously. The model utilizes a shared BERT encoder fine-tuned jointly on both tasks, allowing it to improve generalization from the input text. On top of the shared BERT encoder, we add two task-specific output heads: one for classification and one for regression.

For emotion classification, we use the classification head that consists of a fully connected layer followed by sigmoid activation to produce independent probabilities for each emotion category, enabling multi-label classification. For intensity prediction, we use the regressor head that consists of a fully connected layer followed by a scaled tanh activation to constrain and produce a normalized intensity score between 0 and 1. This architecture allows the model to optimize both tasks in parallel instead of having each task done separately, potentially benefiting from shared representations.

The model is trained on samples that include both an emotion label and an intensity score—unlike the single-task learning approach where each model is trained independently. A joint loss function is used, combining binary cross-entropy loss for multi-label emotion classification and a differentiable Pearson correlation-based loss for intensity prediction. The joint loss function is defined as:

$$L_{total} = \lambda * L_{classification} + (1-\lambda) * L_{regression}$$

where $\lambda \in [0,1]$ controls the weight between cross-entropy loss and Pearson's correlation coefficient. Model training is performed using the standard AdamW optimizer with gradient updates applied to the shared encoder. This shared training setup allows the model to learn shared representation optimized for both classification and regression tasks, potentially improving overall generalization.

C. Dataset

This research utilizes the SemEval-2018 dataset [23], which is a benchmark dataset for affective computing. The dataset consists of English tweets annotated for both emotion classification and intensity prediction, making it suitable for both single-task and multi-task learning approaches.

The dataset consists of multiple subtasks in which we chose one relevant subtask, namely, Emotion Intensity Prediction (El-reg), where tweets are annotated with a real-valued score representing intensity of the emotion expressed, ranging between 0 and 1.

IV. Experiments, Result and analysis

A. Implementation Details

Word Embeddings: We use pre-trained word embeddings available from BERT tokenizer and utilize them to preprocess the input text by tokenizing it into subwords and then converting it into token IDs. The token IDs are then passed into the proposed BERT model to obtain contextualized word embeddings. We find it more suitable over other pre-trained models because our approach is based on BERT architecture, thus ensuring compatibility between the tokenization process and the model's input requirements.

Preprocessing: Before extracting the word embeddings from the input text, we used some preprocessing procedures to clean the data. These include converting all text into lowercase letters, replacing emojis with expression words, expanding contractions, removing URLs present in the text, removing user mentions, removing non-alphabetic characters (except for ! and ?), keeping hashtags and removing excess whitespace.

Synonym Replacement & Noise Injection: A safe synonym replacement and noise injection are applied to create more training sets to balance the number of datasets representing each emotion. Synonym replacement involves selecting random text from the dataset and replacing certain words with their synonyms using the WordNet English Vocabulary [24]. The replacement probability of the word being set as 0.3. We also add noise injection into those new synonym sets to mimic real-life typo. Types of noise injected include skipping random characters and swapping characters. Character skip and swap probability are both set as 0.05. These settings are applied across all our models to ensure the augmented samples are meaningfully distinct from the original dataset.

Lexicons: We incorporate emotion-related features using three types of lexicons, including NRC Emotion Lexicon (EmoLex) [5], NRC Valence-Arousal-Dominance (VAD) Lexicon [25], and NRC Hashtag Emotion Lexicon (HashEmo) [26]

The EmoLex emotion-word association lexicon is used to extract affective features from the text into ten expression vectors: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive, and negative. Each word in the input text is matched against the EmoLex dictionary and the corresponding emotion categories are incremented accordingly.

The NRC VAD Lexicon is used to extract affective features from the text that are grouped into three psychological dimensions: Valence, Arousal, and Dominance. For each word in the text found in the VAD lexicon, its respective scores are accumulated. The final vector representation for the text is computed by averaging the scores across all matched words for each of the three dimensions.

The HashEmo Lexicon is used to extract affective features from the text into 8 basic emotions consisting of: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Each word in the input text is matched against the HashEmo dictionary and the corresponding emotion categories are incremented accordingly. The final vector representation for the text is computed by averaging the scores across all matched words for each of the 8 categories.

All the lexicon features are then scaled using Standard Scaler and then combined together into one vector which consists of 21 dimensions per instance.

Training: We train each model with AdamW optimizer with learning rate of 2e-5 and weight decay of 0.01. Training proceeds for a maximum of 10 epochs using cosine annealing scheduler with warm restarts. Early stopping is applied based on validation performance with a patience of 3 epochs.

For single-task learning, we developed two models: emotion classification model and intensity regression model. The emotion classification model was trained using Cross-Entropy Loss with label smoothing of 0.1 to mitigate confidence, avoid overfitting and improve generalization.

The intensity regression model was trained using a composite loss function that combines the Huber Loss with delta 0.3 and Pearson correlation coefficient to balance between absolute prediction error with correlation trend to the ground truth. The final loss function is defined as:

$$\mathcal{L}\text{total} = \alpha \cdot \mathcal{L}\text{Huber} + (1-\alpha) \cdot \mathcal{L}\text{Pearson}$$

$$\mathcal{L}\text{Pearson} = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} + \varepsilon}$$

where α = 0.3, x and y denote predicted and ground truth intensity values, respectively, and ε = 1e-8 which is a small constant used for numerical stability. The value of α = 0.3 has been chosen to enable more weight (that is, 1 - α = 0.7) to be applied on the Pearson correlation-based loss component. The main reason for this particular choice is that the main evaluation metric on this task, both in the SemEval-2018 dataset and our research, happens to be Pearson's correlation score. In placing more weight on the loss that directly optimizes for correlation, it drives the model to learn and focus more on capturing the relational trend between its predictions and the ground truth values as is desired in intensity regression tasks.

For multi-task learning, we developed 1 multi-task model that could do both emotion classification and intensity regression through shared parameter training. The model was trained using a composite loss function by combining the classification and regression losses.

For the classification loss, we used Binary Cross-Entropy loss. For the regression loss, we used a combination of Huber Loss and Pearson Correlation Loss, identical to the setup of the single-task regression model. The final loss function is defined as:

$$\mathcal{L}\text{total} = \alpha \cdot \mathcal{L}\text{Classification} + (1-\alpha) \cdot \mathcal{L}\text{Regression}, \text{ where } \alpha = 0.3$$

The coefficient α = 0.3 indicates that the intensity regression task is given a higher importance (i.e. 1 - α = 0.7) than the classification task. The reasoning behind this is that the intensity prediction is a more detailed and difficult task in comparison to others, and our multi-task strategy aims at its performance improvement as a major target. By focusing on the regression loss, the common encoder is thus expected to come up with more vivid representations that allow it to interpret the emotional change in the subtle speech, which, indeed, is the main focus of this research.

B. Evaluation

Each classification model is evaluated using precision, recall, f1-score per emotion class and overall accuracy of the model. On the other hand, each regression model is evaluated using Pearson's correlation score. The SemEval2018 El-reg dataset exhibits an imbalanced emotion distribution, with fear being the most represented emotion, followed by anger, joy, and sadness.

1. Classification Model Evaluation

Each classification model is evaluated using precision, recall, f1-score per emotion class and overall accuracy of the model. The SemEval2018 El-reg dataset exhibits an imbalanced emotion distribution, with fear being the most represented emotion, followed by anger, joy, and sadness.

TABLE I. Comparison classification scores

	Single-Task	Multi-Task
joy		
f1-score	.82	.81
precision	.84	.88
recall	.81	.75
sadness		
f1-score	.64	.64
precision	.63	.73
recall	.65	.57
anger		
f1-score	.73	.70
precision	.76	.76
recall	.71	.65
fear		
f1-score	.67	.63
precision	.65	.52
recall	.71	.78
accuracy score	.72	.81

The single-task model shows higher and more consistent performance across individual emotion detection. The emotion joy has the highest performance despite being the second least represented emotion. This is likely because the datasets where emotion is labeled joy are more distinct and less noisy therefore making it easier for the model to identify compared to other emotions. Other emotions like sadness, anger, and fear also show a close range between precision and recall. This stability can be attributed to the single-task model's focused training, which avoids cross-task interference and noise introduced by shared parameters.

This is in contrast with the multi-task model, which jointly performs emotion classification and intensity regression simultaneously through shared parameter training. The emotion fear shows higher recall and lower precision because it's the most represented label, making the model overpredict its emotion and therefore causing more error and less precision in predicting it. In contrast, sadness and anger, which is one of the least represented labels, makes the model more careful when predicting the emotion. Therefore causing their recall to be lower and precision to be higher, because the model only predicts those emotions when it's fairly confident.

While the single-task model often achieves a higher f1-score due to task isolation, the highest overall accuracy score is obtained by the multi-task model. This is because the multi-task model tends to predict emotion dominated by the majority class therefore boosting the overall accuracy even if rare classes are quite poorly predicted. In addition, the shared representation between emotion labels and intensity score might contribute to the final accuracy of multitask learning. Therefore, making the model able to recognize patterns in these emotions, hence the higher accuracy.

2. *Regression Model evaluation*

Evaluation of the regression model uses Pearson's correlation coefficient with a score from 0 to 1.

TABLE II. Comparison Pearson Correlation Scores

Method	a	f	j	s	Avg
Single Task Learning (STL)	.65	.66	.44	.65	.60
Multi Task Learning (MTL)	.73	.75	.72	.72	.73

Single-task learning managed to detect the intensity of fear and anger emotions quite well. This is because both of those emotions are more represented than joy and sadness. Joy shows the lowest pearson correlation score despite not being the least represented value. This is because joy might have a more complex or noisier intensity distribution. The variance value of fear and joy is quite large, which is 0.22. This shows that there is inconsistency in predicting emotion intensity in single-task learning.

This is in contrast to multi-task learning which has a lower variance of 0.03. The highest prediction is fear, followed by anger, joy, and sadness. The order is exactly the same as the order of represented emotion labels in the training sample. Therefore proves that the number of representations in the dataset affect the performance of the multi-task learning model. In addition, multi-task learning shows consistency in predicting emotion intensity, proving that joint learning helps stabilize performance across different emotional dimensions.

The increase in average pearson correlation score from single-task learning (0.60) to multi-task learning (0.73) shows that multi-task learning can not only detect emotion in text, but emotion intensity is also successfully detected quite well. Unlike the classification task benefits from parameter sharing, which allows the model to capture richer emotional representations and leads to more accurate and consistent intensity predictions.

3. *Sample Text Model Evaluation*

TABLE III. Qualitative Sample Texts Comparison

No	Text	Emotion	STL Intensity	MTL Intensity
1	Hey, I just heard back about the project proposal I submitted last week. It got approved, so I'm pretty pleased with how things turned out. It's a good start. #WorkLife	Joy	0.52	0.72
2	OMG I'M LITERALLY SHAKING!!! WE WON THE GRAND PRIZE!!! I CAN'T BELIEVE THIS IS HAPPENING, SO UNBELIEVABLY ECSTATIC RN!!! #BestDayEver #Winner		0.92	0.86
3	I spoke to my grandma earlier, and she's been feeling a bit lonely lately. It makes me feel a bit down just thinking about her being by herself so much.	Sadness	0.73	0.72
4	I just got back from visiting my grandma, and seeing her so sad over her passing husband... it just completely shattered me. My heart is just broken thinking about it.		0.91	0.87
5	You know, I'm trying to work here, but my coworker keeps interrupting me with questions he could easily find the answers to himself. Honestly, it's starting to get really annoying.	Anger	0.57	0.54
6	I cannot believe what my coworker just did! He went to our boss and took credit for the entire project I've been working on for months. I'm absolutely furious and shaking with rage right now!		0.92	0.86
7	Just experienced an earthquake tremor a few minutes prior. The building was shaking for a bit, my heart is still racing though.	Fear	0.81	0.75
8	Have to give a big presentation in a few minutes. Feeling the nerves kicking in now. #PublicSpeaking		0.83	0.91

The table above shows the qualitative comparison of intensity prediction from the STL and MTL models. Both models can distinguish emotion trends between low level intensity and high level intensity texts. The STL model shows a strong capability to distinguish intensity for emotions like *joy* and *anger* with high variance by struggles with *sadness* and *anger* showing lower variance. In contrast, the MTL model demonstrates a more consistent and balanced performance across all emotions, making it a more effective and robust model overall.

V. Conclusion

In this work, we have proposed a single-task learning and multi-task learning model to do two tasks: emotion classification and intensity prediction of a text. We evaluated the classification using f1-score, precision, and recall. Meanwhile, the regression model was evaluated using Pearson's correlation coefficient. Experiment results show that the multi-task model performs better than single-task model. This is because multi-task learning models are trained with shared parameters which increases the knowledge of the models. While it might increase noise and task-interference, our experiment shows that the multi-task model still performs better than the single-task model. The difference is especially shown in the increase of the intensity regression score.

Acknowledgement

The authors would like to express their deepest gratitude to the Faculty of Computer Science, Bina Nusantara University, for the support and guidance provided during the making of this research. This research would not have been possible without the academic environment and encouragement provided by the supervisors. We appreciate the opportunity and knowledge that has contributed greatly to the completion of this research.

References

[1] A. Devgan, "Contextual Emotion Recognition Using Transformer-Based Models," *Authorea Prepr.*, 2023, Accessed: Apr. 19, 2025. [Online]. Available: https://www.academia.edu/download/105292021/IRJET_V10I7150.pdf

[2] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.

[3] S. Devaram, "Empathic chatbot: Emotional intelligence for mental health well-being," *ArXiv Abs201209130*, 2020.

[4] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances," May 08, 2019, *arXiv*: arXiv:1905.02947. doi: 10.48550/arXiv.1905.02947.

[5] S. M. Mohammad and P. D. Turney, "CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, Aug. 2013, doi: 10.1111/j.1467-8640.2012.00460.x.

[6] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 579–586. Accessed: Apr. 19, 2025. [Online]. Available: https://aclanthology.org/H05-1073.pdf

[7] S. M. Mohammad and F. Bravo-Marquez, "Emotion Intensities in Tweets," Aug. 11, 2017, *arXiv*: arXiv:1708.03696. doi: 10.48550/arXiv.1708.03696.

[8] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017, doi: 10.1016/j.inffus.2017.02.003.

[9] Y. Wang *et al.*, "A systematic review on affective computing: emotion models, databases, and recent advances," *Inf. Fusion*, vol. 83–84, pp. 19–52, Jul. 2022, doi: 10.1016/j.inffus.2022.03.009.

[10] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Nat. Lang. Eng.*, vol. 23, no. 5, pp. 649–685, Sep. 2017, doi: 10.1017/S1351324916000383.

[11] S. Buechel and U. Hahn, "EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis," May 04, 2022, *arXiv*: arXiv:2205.01996. doi: 10.48550/arXiv.2205.01996.

[12] S. Mohammad, "A Practical Guide to Sentiment Annotation: Challenges and Solutions," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, A. Balahur, E. van der Goot, P. Vossen, and A. Montoyo, Eds., San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 174–179. doi: 10.18653/v1/W16-0429.

[13] S. M. Mohammad and F. Bravo-Marquez, "WASSA-2017 Shared Task on Emotion Intensity," Aug. 11, 2017, *arXiv*: arXiv:1708.03700. doi: 10.48550/arXiv.1708.03700.

[14] M. Köper, E. Kim, and R. Klinger, "IMS at EmoInt-2017: Emotion Intensity Prediction with Affective Norms, Automatically Extended Resources and Deep Learning," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, A. Balahur, S. M. Mohammad, and E. van der Goot, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 50–57. doi: 10.18653/v1/W17-5206.

[15] D. Kulshreshtha, P. Goel, and A. Kumar Singh, "How emotional are you? Neural Architectures for Emotion Intensity Prediction in Microblogs," in *Proceedings of the 27th International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2914–2926. Accessed: Apr. 20, 2025. [Online]. Available: https://aclanthology.org/C18-1247/

[16] M. S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "A Multi-task Ensemble Framework for Emotion, Sentiment and Intensity Prediction," Oct. 15, 2018, *arXiv*: arXiv:1808.01216. doi: 10.48550/arXiv.1808.01216.

[17] M. Hasan, E. Rundensteiner, and E. Agu, "DeepEmotex: Classifying Emotion in Text Messages using Deep Transfer Learning," Jun. 12, 2022, *arXiv*: arXiv:2206.06775. doi: 10.48550/arXiv.2206.06775.

[18] F. Mohammad *et al.*, "Text Augmentation-Based Model for Emotion Recognition Using Transformers," *Comput. Mater. Contin.*, vol. 76, no. 3, pp. 3523–3547, 2023, doi: 10.32604/cmc.2023.040202.

[19] I. N. Alejo, T. Badia, and J. Barnes, "Cross-lingual Emotion Intensity Prediction," Nov. 24, 2020, *arXiv*: arXiv:2004.04103. doi: 10.48550/arXiv.2004.04103.

[20] A. S. Sundari and R. Shenbagavali, "Dominant Lexicon Based Bi-LSTM for Emotion Prediction on a Text," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 11S, pp. 1272–1277, Oct. 2019, doi: 10.35940/ijitee.K1256.09811S19.

[21] Z. Teng, D.-T. Vo, and Y. Zhang, "Context-Sensitive Lexicon Features for Neural Sentiment Analysis," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds., Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1629–1638. doi: 10.18653/v1/D16-1169.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[23] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1–17. doi: 10.18653/v1/S18-1001.

[24] G. A. Miller, "WordNet: A Lexical Database for English," in *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. Accessed: May 24, 2025. [Online]. Available: https://aclanthology.org/H94-1111/

[25] S. Mohammad, "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words," in *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), I. Gurevych and Y. Miyao, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 174–184. doi: 10.18653/v1/P18-1017.
[26] S. M. Mohammad and S. Kiritchenko, “Using Hashtags to Capture Fine Emotion Categories from Tweets”.