**Project Summary**

**Emotion Classification using Tweets**

Xuda Chen

Phakamin Tantiwatanavorakul

Allen Zhang

Nam Ken Yan Bouchara

Kang Zhu

1. **Introduction**

This study focuses on emotion classification in tweets, a significant task in Natural Language Processing (NLP). The goal of our project is to perform text classification using advanced artificial intelligence algorithms to accurately label tweets with the emotions expressed by their authors, aiming to achieve increased performance on relevant metrics. Our research investigates and compares various approaches to this task. We evaluate the effectiveness of traditional machine learning methods, such as Logistic Regression and Naïve Bayes, against more advanced deep learning architectures. These include Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), transformer-based models like BERT, and a more complex model (BERT+BILSTM). By implementing and assessing these diverse AI techniques, we aim to identify the most effective methods for correctly labeling the emotional content of tweets.

1. **Related Work**

Emotion classification has been approached using various techniques, particularly in the domain of textual classification. Abdul-Mageed and Ungar (2017) demonstrated the effectiveness of Gated Recurrent Neural Networks (GRNN) for fine-grained emotion detection. Their work highlights the potential of deep learning models in capturing nuanced emotional information from text.

Bharti et al. (2022) utilized deep learning models, including Gated Recurrent Units (GRU) and Bidirectional GRU (Bi-GRU), combined with embeddings such as Word2Vec, GloVe, ELMo, and FastText. They applied these techniques to datasets like ISEAR, WASSA, and Emotion-Stimulus, achieving notable results. Their Bi-GRU model reached an accuracy of 79.46%, while their hybrid model achieved a precision of 82.39%, a recall of 80.40%, and an F1-score of 81.27%.

Zhou and Wu (2018) explored Long Short-Term Memory (LSTM) networks and Bidirectional LSTM (Bi-LSTM) with attention mechanisms. Using the WASSA2018 Implicit Emotion Shared Task dataset, they achieved F1-scores of 0.685 for both LSTM and Bi-LSTM, demonstrating the effectiveness of these models for implicit emotion classification.

Saravia et al. (2018) introduced CARER (Contextualized Affect Representations for Emotion Recognition), which integrates multi-layer CNN architecture with enriched patterns for emotion recognition. They used English tweets from the Twitter API, preprocessing the data through normalization, graph aggregation, token categorization, and pattern extraction. Their CARER model achieved the highest average F1-score of 0.79, significantly outperforming other models such as CNNs, RNNs, and Bi-GRNNs.

Devlin et al. (2018) presented BERT, a pre-trained deep bidirectional transformer model that has become foundational in many NLP tasks. BERT's architecture allows for comprehensive context understanding, which is particularly useful for emotion classification tasks.

In summary, the combination of these approaches—GRNN, Bi-GRU, LSTM, Bi-LSTM, CARER, and BERT—demonstrates the diverse methodologies employed in emotion classification. Each method contributes uniquely to improving performance in detecting and classifying emotions from textual data. We also attempt to fine-tune these models on specific datasets to effectively address the emotion classification task.

1. **Methods**

We chose Word2Vec as the embedding method for most of our models, except for the BERT and BERT-BiLSTM models, which use the embedding layers in BERT. Because Word2Vec is good at capturing local contextual relationships, especially in short-term dependencies, allowing the model to understand context better.

The model that is used in this experiment can be separated into 3 categories: machine learning, deep learning, and transformer:

In machine learning, we use logistic regression and naive bayes which are very popular tools for classifying tasks.

The deep learning part contains the LSTM and GRU also the bi-directional versions (BiLSTM and BiGRU). These models can study the context and perform well in many NLP applications.

The transformer part, we have the BERT model developed by Google which is proposed to be the state-of-the-art of the NLP model with the help of transformer attention characteristic. Difference from the LSTM and GRU which read each word one by one, this model reading the entire sequence at once. Specifically we use BERT-base model which is a pre-trained model but not fine-tuning. We also use BERT-BiLSTM which combines BERT with a bidirectional Long Short-Term Memory network (BiLSTM), further improving the model's ability to handle long-range dependencies.

1. **Experimental Setup**

The dataset used in this project is sourced from the EMOTION dataset available at <https://huggingface.co/datasets/dair-ai/emotion> . It consists of 416,809 texts with associated labels. We divided this unsplit dataset into training, testing, and validation sets with the following proportions: 80% for training, 10% for testing, and 10% for validation. The dataset contains six emotion categories: sadness, joy, love, anger, fear, and surprise. These labels are imbalanced. To address this, we resampled the texts in the training set to ensure balance, resulting in 11,929 samples for each emotion category. Additionally, we removed punctuation and stop words from the texts, as these elements are considered noise for the classification task.

There are some other findings based on the data exploration, such as ‘feel’ appears in 99% sentences of the dataset; the sentences with no obvious emotions will be labeled as ‘love’; The sentences with a negative tone are incorrectly labeled.

In this project we are using the same word embedding method for all of the designed models except BERT and BERT BiLSTM.

The detail (hyperparameters) of doing the models are:

Logistic regression is the machine learning model used for classifying the difference classes. This model is build from the linear layer with the 300\*35 input neurons and 6 output neurons and AdamW optimizer with cross entropy loss function.

Naive Bayes model we use from the scikit-learn pre-trained model (https://scikit-learn.org/stable/modules/naive\_bayes.html ) to do the classification task.

The deep learning models (LSTM, BI-LSTM, GRU, BI-GRU) have the same hyperparameters and the output layers. All of them have softmax classification output, cross entropy loss function, and AdamW optimizer. Learning rate of 0.01, dropout layer for LSTM or GRU of 0.1, dropout layer for output layer 0.1, number of LSTM or GRU layers are 2, embedding dimension of 300, hidden dimension of 128, and output dimension of 6.

The pre-trained BERT models we use are from the hugging face library which can be found here <https://huggingface.co/docs/transformers/model_doc/bert.> It employs 12 layers of transformer blocks, a hidden size of 768, 12 self-attention heads, and approximately 110 million trainable parameters. On top of this model, we have added a fully connected layer with 6 output units and trained it for classification tasks. Importantly, we did not freeze any of the model’s parameters during training, allowing the BERTBase model’s weights to be updated alongside the newly added fully connected layer.The learning rate is 2×10^(-5) and dropout output layer of 0.1 with the output dimension of 6. The optimizer and the loss function and the output function are the same as deep learning.

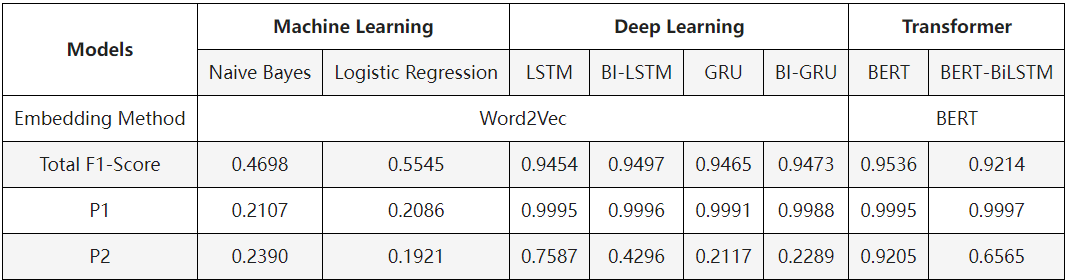
The BERT-BiLSTM is the modification of the BERT model so we are using the same BERT model with the BiLSTM and the classification layer. The hyperparameters are learning rate 2×10^(-4), dropout in LSTM and output layer of 0.8 and 0.5 respectively. 2 layers of BiLSTM with hidden dimension of 128 and output dimension of 6.

For evaluation metrics, we compare several key performance indicators, with a focus on the F1-score. because the F1-score provides a single metric that incorporates both precision and recall, allowing for a more comprehensive evaluation of model performance than using accuracy alone. Additionally, we use our models to predict some specific sentences to evaluate their performance, particularly those sentences containing the word "feel" and those without "feel". Because the word 'feel' appears in 99% of the sentences in the dataset, some models may overly rely on this word for making predictions.

1. **Results**

The table below shows the results of our experiment:

Table 1. Performance of Various Models on Classification Task.



Note:

1. 'P1' refers to the probability of the emotion 'joy' when predicting the sentence 'I feel happy'

2. 'P2' refers to the probability of the emotion 'joy' when predicting the sentence 'I become happy'

From this table we can see that:

BERT has the highest F1-score among all models, and it performs significantly better than other models in predicting the emotion of sentences that do not contain the word 'feel'. It is the only model with a predicting probability of the correct emotion exceeding 0.9. The reason is BERT's deep contextual understanding and attention mechanisms allow it to effectively capture the nuanced emotional content of sentences, making it robust against specific word biases.

Deep learning models, including Transformer-based models, have significantly higher F1-scores compared to machine learning models. All deep learning models achieve an F1-score above 0.9, whereas machine learning models fall below 0.6. Because deep learning models can automatically learn and extract complex features from raw data, which allows them to capture subtle patterns and nuances in emotions that traditional machine learning models might miss.

The total f1-score of BI-LSTM, BI-GRU is larger than GRU, LSTM. Because they are bidirectional, allowing them to capture information from both past and future contexts in a sequence. This bidirectional nature helps in understanding the context more comprehensively, leading to better performance.

BERT-BiLSTM has lower F1-score than other deep learning models. This may be because the model is too complex, while the dataset is too simple, leading to overfitting

1. **Conclusions**

We can conclude that the BERT model is performing the best thanks to BERT's deep contextual understanding and attention mechanisms which allow it to effectively capture the nuanced emotional content of sentences, making it robust against specific word biases.

Additionally, deep learning performs better than machine learning; the bidirectional models have better accuracy than the unidirectional version.

There are also some limitations:

BERT-BiLSTM is performing worse than deep learning which is not what we expect. The reason come from the complexity of the model and the dataset is not rich enough for the model, which lead to overfitting. It highlights the need for balancing model complexity with dataset complexity to achieve optimal performance.

Additionally, some models are overly rely on the word ‘feel’ for making predictions; some sentences with negation words are not correctly labeled.

To address the current limitations, future research should focus on developing or obtaining a richer dataset that goes beyond the simplistic pattern of "feel" + emotion words. A more sophisticated dataset with a complex structure is needed to accurately capture and label sentiments, including those expressed in sentences with negation.

Reference

Abdul-Mageed, M., & Ungar, L. (2017). EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 718–728).

Bharti, S. K., Varadhaganapathy, S., Gupta, R. K., Shukla, P. K., Bouye, M., Hingaa, S. K., & Mahmoud, A. (2022). Text-based emotion recognition using Deep Learning Approach. *Computational Intelligence and Neuroscience, 2022*, 1–8. <https://doi.org/10.1155/2022/2645381>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. <https://doi.org/10.48550/arXiv.1810.04805>

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018). Carer: Contextualized affect representations for emotion recognition. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.18653/v1/d18-1404>

Zhou, Q., & Wu, H. (2018). NLP at IEST 2018: BiLSTM-attention and LSTM-attention via soft voting in emotion classification. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. <https://doi.org/10.18653/v1/w18-6226>