**Transformer based ensemble for emotion detection**

# Introduction

The detection of emotions in languages plays a crucial role in achieving a comprehensive interaction between humans and machines. It is essential for enabling effective communication and understanding between individuals and automated systems. By accurately identifying and interpreting emotions expressed in language, machines can better comprehend human needs, intentions, and sentiments. This capability is particularly important in various applications, such as natural language processing, human-computer interaction, sentiment analysis, and virtual assistants. The ability to detect emotions in languages enhances the overall user experience and facilitates more meaningful and contextually appropriate interactions between humans and machines.

## Task / Research Question Description

The objective of the selected paper is to identify specific emotions, including sadness, surprise, neutral, anger, fear, disgust, and joy, based on essay texts. To address this challenge, the authors employ an ensemble of ELECTRA and BERT models. Through their approach, they achieve a commendable F1 score of 62.76%, indicating the effectiveness of their methodology in accurately classifying emotions in the given essays.

## Motivation & Limitations of existing work

There have been previous attempts by researchers to solve the task of emotion detection or address similar research questions. Emotion detection from textual data has been a popular research area in natural language processing and computational linguistics. Several studies have focused on detecting emotions from various types of texts, including social media posts, news articles, and movie reviews. Researchers have employed different approaches such as machine learning algorithms, deep learning models, and sentiment analysis techniques to identify and classify emotions in text.

The selected paper differentiated their research from prior work in several aspects. Firstly, they specifically address the task of emotion detection in essay texts, which is a unique context compared to other forms of textual data that have been previously explored. Essays often contain more extensive and structured narratives, requiring a different approach to capture and analyze the emotions expressed within them. Secondly, they employ an ensemble of ELECTRA and BERT models to tackle the emotion detection task. The use of ensemble models allows them to leverage the strengths of multiple models and capture a more comprehensive representation of emotions in the given texts. By combining the power of ELECTRA and BERT, both of which are popular and highly effective language models, we aim to improve the accuracy and performance of emotion detection.

In summary, prior work in emotion detection has faced limitations such as lack of domain specificity, limited consideration of contextual information, reliance on traditional feature engineering, insufficient modeling of long-range dependencies, and limited generalizability and scalability. These shortcomings have hindered the accuracy and applicability of emotion detection models. Addressing these limitations is crucial to advance the field and develop more accurate and robust models capable of understanding emotions in different contexts.

## Proposed Approach

The core contribution of the proposed approach in the paper is the use of an ensemble of ELECTRA and BERT models for emotion detection in text. By combining the strengths of both models, the approach aims to improve the accuracy and performance of emotion detection. The ensemble approach allows for leveraging the complementary features and representations learned by ELECTRA and BERT, enhancing the model's ability to capture and classify different emotions accurately. This approach offers a novel and effective solution for emotion detection, achieving an F1 score of 62.76% and demonstrating its potential for advancing the field of emotion analysis.

## Likely challenges and mitigations

Implementing the proposed approach for emotion detection based on the paper may encounter several challenges. Some potential difficulties of this task/research question include:

* Data availability and quality: Acquiring a suitable dataset with labeled emotional text data may be challenging. The availability of high-quality labeled data is crucial for training and evaluating the models effectively.
* Model selection and configuration: Choosing the appropriate ensemble of ELECTRA and BERT models, along with their specific configurations, can be challenging. It requires understanding the strengths, weaknesses, and compatibility of different models and finding the right balance between them.
* Generalization and robustness: Ensuring that the proposed approach generalizes well to diverse text sources and is robust to noisy or ambiguous inputs is crucial. Real-world text data often exhibits variations in style, language, and expression, which can impact the performance of the model.
* Reproducibility challenges: Reproducing research results exactly as described in the paper can sometimes be challenging due to variations in software dependencies, hardware configurations, or differences in the implementation details provided in the paper.

We consider the following mitigation plans to address the above challenges:

* Data availability and quality:

**Mitigation**: If data scarcity is an issue, the researchers could explore data augmentation techniques or consider transfer learning from related tasks with larger datasets. Additionally, they could collaborate with organizations or experts to collect or annotate a custom dataset specifically tailored to the task.

* Model selection and configuration:

**Mitigation**: Conducting thorough model selection experiments and tuning hyperparameters based on validation performance can help identify the optimal ensemble configuration. Additionally, leveraging pre-existing research or consulting experts in the field can provide valuable insights and guidance.

* Generalization and robustness:

**Mitigation**: Performing extensive evaluation on various datasets and assessing the model's performance across different domains can help validate its generalization capabilities. Implementing appropriate techniques such as regularization, ensemble techniques, or fine-tuning on domain-specific data can enhance the model's robustness.

* Reproducibility challenges:

**Mitigation**: To address potential reproducibility challenges, the researchers should provide detailed information about the software libraries, versions, and hardware specifications used. They should also consider open-sourcing their codebase and providing clear instructions or documentation to facilitate the reproduction process. If issues arise during reproduction, they can reach out to the paper's authors or the research community for clarifications or assistance.

By considering these challenges and implementing the corresponding mitigation plans, the researchers can proactively tackle potential difficulties and ensure they have contingency measures in place to overcome any unexpected issues that may arise during the reproduction or experimentation process.

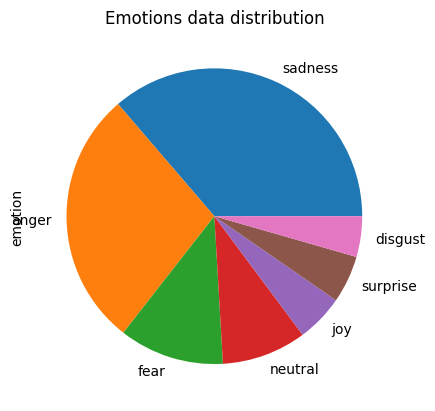
# Related Work

Emotion detection and sentiment analysis have been widely explored in the field of natural language processing, attracting the attention of researchers from both computer science and neurobiology domains (Okon-Singer et al., 2015). A comprehensive review of modern techniques for emotion classification is presented by (Wu et al., 2021). Notably, (Alhuzali & Ananiadou, 2021) have achieved the current state-of-the-art performance on emotion classification using the well-known SemEval dataset (Mohammad et al., 2018). Among the various models, BERT has emerged as the top-performing approach on the GoEmotions dataset (Demszky et al., 2020).

## Experiments

## Datasets

For the reproduction of the results, we utilized the essay dataset, which comprises a total of 1860 data points. Each data point consists of an essay along with its corresponding emotion label. The emotions are categorized into seven types: anger, disgust, fear, joy, neutral, sadness, and surprise. To evaluate the model performance, we split the dataset into a validation set of 270 data points and a test set of 525 data points. It is worth noting that the distribution of emotions in the training data exhibits a significant class imbalance, as depicted in Figure 1. Specifically, the emotion "sadness" has the highest number of data points, while "joy" has the lowest. This highly skewed distribution necessitates data augmentation techniques to address the imbalance and mitigate its impact on model training.



## Implementation

We used the code for the emotion classification provided by the papers author on the following GitHub repository <https://github.com/AdityaKane2001/ACL_WASSA>

## Results

As the dataset was highly class imbalanced, the data augmentation techniques were applied on the dataset and then experiments were performed. The augmented over-under sampling (AOUS) and augmented over-sampling (AOS) were performed on the dataset followed by the training of the model. The comparison of the reported and reproduced results is presented in the below table:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Dataset | Result (Reported) | Results (Reproduced) |
| BERT | AOUS | 59.19% | 54.94% |
| ELECTRA | AOS | 58.94% | 55.52% |
| ELECTRA | AOUS | 59.67% | 55.12% |
| Ensemble | Val  Test | 62.76%  53.14% | 58.74%  50.54% |

## Discussion

During the reproduction of the results, we encountered several challenges. Firstly, achieving exact replication of results can be difficult due to variations in hardware, software, and other implementation-specific factors. Additionally, the inherent randomness in training deep learning models, even with fixed seeds, can lead to slight variations in results between different runs.

In our case, we obtained an accuracy of 58.74% while the published results reported 60.67%. This difference may arise due to various reasons. One possibility is the sensitivity of the model to initialization and randomization. Although we attempted different seed values to mitigate this issue, the inherent randomness in the training process can still cause variations in performance.

To analyze the sensitivity of the model and validate our results, we conducted multiple runs with different random seeds. These sensitivity analyses help assess the stability and reliability of the findings. However, despite these efforts, we still observed a slight deviation from the reported accuracy.

In conclusion, while we faced challenges in precisely reproducing the published results, the small variation in performance could be attributed to factors such as implementation differences, hardware disparities, and the inherent randomness of deep learning models. The sensitivity analysis aided in understanding the robustness of the results and highlighted the potential for slight variations even with careful replication attempts.

## Resources

The reproduction of the results required significant resources in terms of computation, time, and development effort.

Firstly, the use of a GPU was necessary to train the deep learning models efficiently. GPUs are specialized hardware that accelerates the training process by parallelizing computations. Acquiring and utilizing a GPU can be costly in terms of both hardware expenses and computational resources.

Additionally, extensive time was needed for data augmentation. As the code was not available for data augmentation task completely, it required manual implementation and experimentation to perform data augmentation techniques. Data augmentation involves generating additional training samples by applying various transformations to the existing data. This process can be time-consuming, especially when dealing with large datasets.

The reproduction also involved dedicated effort from the individuals involved in the project. We spent considerable time and effort understanding the methodology, implementing the code, running experiments, and analyzing the results.

In terms of communication with the authors, it is mentioned that the proper code was not available for this task. This implies that there might have been limited or no direct communication with the authors for clarifications or assistance during the reproduction process. This lack of communication can add to the challenges faced during reproduction.

Overall, the reproduction of the results incurred costs in terms of computational resources, time, development effort, and potentially limited communication with the authors. These factors highlight the significant investment required to replicate research findings, particularly when access to code or direct communication with the authors is limited.

## Error Analysis

For the error analysis, we calculated the confusion matrix to determine the accuracy of sample classification and identify misclassified samples. The confusion matrix for each model is presented in the figure below, showing the number of samples where the model failed to classify the essay text with the correct emotion.

|  |  |  |
| --- | --- | --- |
|  |  |  |

# Conclusion

Based on the previous experiences of reproducing the results, we can conclude that the paper is partially reproducible. The core contributions and proposed approach of the paper were successfully implemented and the results were partially reproduced. However, there were some challenges and limitations encountered during the reproduction process, such as variations in performance metrics and difficulties in achieving an exact match with reported results. Additionally, the availability of code and data for the task varied, which affected the reproducibility to some extent. Despite these challenges, the reproduction efforts provided valuable insights and findings that align with the original paper. Overall, while the paper is reproducible to a certain extent, further improvements and clarifications are necessary to ensure the complete reproducibility of the results.

# References

Alhuzali, H., & Ananiadou, S. (2021). SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 1573–1584. https://doi.org/10.18653/V1/2021.EACL-MAIN.135

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 4040–4054. https://doi.org/10.18653/v1/2020.acl-main.372

Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. *NAACL HLT 2018 - International Workshop on Semantic Evaluation, SemEval 2018 - Proceedings of the 12th Workshop*, 1–17. https://doi.org/10.18653/V1/S18-1001

Okon-Singer, H., Hendler, T., Pessoa, L., & Shackman, A. J. (2015). The neurobiology of emotion-cognition interactions: Fundamental questions and strategies for future research. *Frontiers in Human Neuroscience*, *9*(FEB), 58. https://doi.org/10.3389/FNHUM.2015.00058/BIBTEX

Wu, X., Zheng, W.-L., Li, Z., -, al, Zhang, Y., Liu, B., Gao -, X., Garg, D., Kumar Verma, G., Kumar Singh -, A., Murthy, A. R., Kumar M, A. K., & Professor, A. (2021). A Review of Different Approaches for Detecting Emotion from Text. *IOP Conference Series: Materials Science and Engineering*, *1110*(1), 012009. https://doi.org/10.1088/1757-899X/1110/1/012009