Comparative Study of Transformers Models in Understanding Emotions

Heaven Klair
heavenklair@berkeley.edu

Abstract

This paper delves into the nuanced task of emotion classification in online communication, focusing on analyzing Reddit posts. The core objective is to accurately categorize each post into one of 28 distinct categories: 27 specific emotions or a Neutral state. This endeavor leverages a rich dataset of approximately 43,000 instances, marking a significant intersection between Natural Language Processing (NLP) and emotional analysis in digital interactions. This project explored the use of transformer models, chosen for their proven efficacy in capturing the subtle complexities and contextual nuances inherent in textual emotional expression.

1 Introduction

The rise of online platforms like Reddit has generated an unprecedented amount of text data rich in emotional content. This paper is motivated by the challenge of understanding and categorizing these emotions effectively. Accurately identifying and classifying the emotional tone of posts can significantly aid in analyzing online interactions. Particularly, it can enable platforms to proactively flag and address hateful content, contributing to healthier online environments.

Transformers are renowned for their ability to process text while capturing contextual nuances, making them ideal for emotion classification. **Traditional linear text processing models struggle with the complexity and subtlety required for this task. By contrast, transformers employ attention mechanisms to emphasize the significance of different words in a sentence.** This paper explores the application of transformers to classify emotions in Reddit posts, a choice driven by their effectiveness in handling diverse and nuanced textual data.

The study's baseline is established using a Naive Bayes classifier, achieving an F1-macro score of 0.06 on the dataset. This score serves as a benchmark, highlighting the limitations of simpler NLP models in capturing the intricacies of emotional expressions in Reddit posts. The performance gap between the Naive Bayes model and advanced transformer models, as explored in this paper, will offer insights into the progress and applicability of modern NLP techniques for real-world challenges.

2 Background

A cornerstone in the domain of emotion analysis in text is the "GoEmotions: A Dataset of Fine-Grained Emotions" (1) study, which introduced the GoEmotions dataset. It leveraged the Bidirectional Encoder Representations from Transformers (BERT) model (3) to validate the robustness of the dataset. The BERT-based approach achieves an average F1-score of 0.46 across the proposed taxonomy, highlighting both the effectiveness of BERT in emotion classification and the potential for further advancements in this field.

Example Posts	Associated Labels
We need more boards and to create a bit more space for [NAME]. Then we'll be good	desire, optimism
maybe that is what happened to the great white at houston zoo	confusion, realization
hello everyone. i am from toronto as well. can call and visit in personal if needed	neutral
demographics? i do not know anybody under 35 who has cable tv.	confusion
aww... she will probably come around eventually, i am sure she was just jealous of [name]... i mean, what woman would not be! lol	amusement, approval

Table 1: Some Example from the GoEmotions Dataset

The study Seq2Emo: A Sequence to Multi-Label Emotion Classification Model (2) introduces a sequence-to-emotion approach, utilizing a bi-directional decoder in an Encoder-Decoder framework with LSTM networks. This innovative model,

when tested on the SemEval'18 and GoEmotions datasets, outperforms existing methods like Binary Relevance and Classifier Chain approaches, showcasing its ability to effectively capture complex emotional correlations. Complementing this, "Uncovering the Limits of Text-based Emotion Detection"(4) explores the challenges in text-based emotion detection through a multi-label classification strategy focused on minimizing binary cross-entropy loss over various target emotions. The methodologies employed, including Deep Neural Networks and Bi-directional LSTMs, provide crucial insights into the strengths and limitations of current emotion detection techniques, enhancing the understanding of emotion classification in NLP.

3 Experimental Setup

This section outlines the methodology employed in this study, focusing on the implementation and evaluation of various transformer-based models for emotion classification in Reddit posts. The goal is to leverage these models to accurately classify text into one of the 27 specific emotions or a Neutral state, as defined in the GoEmotions dataset. All the transfomer models used word embeddings of length 128. For each of the following transformer models, the entire BERT was set as trainable during the training proccess.

1. BERT-base Cased Model: The study first experimented with the deployment of an out-of-the-box BERT-base cased model. This model was specifically chosen for its bidirectional training, which allows it to understand the context of a word based on all of its surroundings (both left and right of the word). The BERT-base cased model achieved a Macro F1-score of 0.46 and a Macro precision of 0.58.

2. DistilBERT (5): Following the BERT model, the study experimented with DistilBERT, a streamlined version of the original BERT model. DistilBERT retains most of the original model's performance capabilities but is smaller and faster, making it more efficient for practical applications. This model's implementation aimed to examine whether similar performance levels could be achieved with reduced computational resources.

3. RoBERTa: This model modifies key hyperparameters in BERT, removing the Next Sentence Prediction objective and training with much larger mini-batches and learning rates. The implementation of RoBERTa aimed to assess whether these modifications could enhance performance in emotion classification tasks compared to the standard BERT model.

4. XLNet (6): Finally, the study incorporated XLNet, a generalized autoregressive pretraining method. XLNet differs from BERT in that it captures bidirectional context by using a permutation-based training approach. This implementation was aimed at exploring whether XLNet's unique training methodology could further improve the classification of emotions in textual data.

Model	Precision	Recall	F1-Score
Naive Bayes Classifier	0.25	0.04	0.06
BERT-base Cased	0.54	0.41	0.45
DistilBERT	0.62	0.39	0.45
Roberta BERT	0.65	0.35	0.42
XLNET	0.62	0.35	0.41

Table 2: Performance Metrics of Different Models

Each of these models was selected based on their individual strengths in understanding and classifying emotions content in a text. The comparative analysis of these models' performance on the GoEmotions dataset provides a comprehensive understanding of their applicability and effectiveness in the domain of NLP and emotion classification.

4 Results

The models were evaluated based on key metrics: Precision, Recall, and F1-Score. These metrics provide a comprehensive understanding of each model's ability to accurately classify text into the defined emotional categories. The performance of each model, as shown in the table 1, offers insights into their relative effectiveness and efficiency.

The Naive Bayes Classifier, the baseline model, demonstrates limited performance with an F1-Score of 0.06. This highlights the inherent complexities of emotion classification and the inadequacy of simpler models for this task.

The BERT-base cased model and the DistilBERT model showed a notable improvement with both an F1-Scores of 0.45. This would set a benchmark for comparing the performance of the other models implemented in this study. DistilBERT got a higher precision than BERT-base Cased model. It attained Precision of 0.62 whereas BERT-base Cased got an

	precision	recall	f1-score	support
admiration	0.69	0.62	0.65	504
amusement	0.79	0.81	0.80	264
anger	0.62	0.30	0.41	198
annoyance	0.50	0.18	0.26	320
approval	0.45	0.34	0.39	351
caring	0.41	0.33	0.37	135
confusion	0.35	0.37	0.36	153
curiosity	0.51	0.32	0.40	284
desire	0.56	0.34	0.42	83
disappointment	0.32	0.23	0.27	151
disapproval	0.48	0.28	0.36	267
disgust	0.64	0.33	0.44	123
embarrassment	0.67	0.32	0.44	37
excitement	0.55	0.31	0.40	103
fear	0.65	0.73	0.69	78
gratitude	0.90	0.91	0.91	352
grief	0.00	0.00	0.00	6
joy	0.58	0.55	0.56	161
love	0.74	0.84	0.78	238
nervousness	0.38	0.22	0.28	23
optimism	0.57	0.51	0.54	186
pride	0.60	0.19	0.29	16
realization	0.48	0.16	0.24	145
relief	0.25	0.09	0.13	11
remorse	0.60	0.55	0.57	56
sadness	0.54	0.53	0.54	156
surprise	0.54	0.56	0.55	141
neutral	0.68	0.54	0.60	1787
micro avg	0.63	0.49	0.55	6329
macro avg	0.54	0.41	0.45	6329
weighted avg	0.61	0.49	0.54	6329
samples avg	0.54	0.51	0.51	6329

Figure 1: Performance Metrics of Individual Emotion on BERT-base Cased

precision of 0.54. XLNET and RoBERTa models acheived F1-score of 0.42 and 0.41 respectively.

5 Discussion

The figure 1 below displays the performance report of individual emotions from the BERT-base Cased Model. Among the evaluated models, BERT-base Cased and DistilBERT were the top performers, particularly in terms of their Macro F-1 scores. A closer look at the performance report for each emotion reveals that the dataset includes merely six instances of *grief*.

Significantly, for this emotion, the metrics of precision, recall, and F1-score all came out to be zero. Additionally, the emotion *relief* only occurs 11 times in the test set, while *pride* accounts for 16 instances. The limited representation of these emotions in the dataset is directly linked to their lower performance metrics.

Classifying emotions into 28 distinct labels is challenging due to the nuanced and overlapping nature of emotional states. Many emotions are context-dependent and can manifest similarly in text, making it difficult for models to distinguish between them with high precision and recall.

Adopting the Ekman emotion wheel to group 27 emotions from GoEmotions into 7 broader categories resulted in a notable improvement in model's performance. This grouping strategy aligns with the psychological theory that suggests a smaller number of primary and universal emotions. The emotions were grouped as follows: *anger, annoyance, disapproval* under *anger*, *disgust* remained as itself, *joy*, *amusement, approval, excitement, gratitude, love, optimism, relief, pride, admiration, desire, caring*, *sadness, disappointment, embarrassment, grief, remorse* under *sadness*, *surprise, realization, confusion, curiosity* under *surprise*, and *neutral* is remained by itself.

By doing so, the model learns more effectively and differentiate between the broader emotional categories. This grouping reduces the complexity of the classification task and diminishes the issues arising from the sparse representation of certain emotions in the dataset.

	precision	recall	f1-score	support
anger	0.57	0.47	0.51	726
disgust	0.63	0.40	0.49	123
fear	0.67	0.51	0.58	98
joy	0.80	0.82	0.81	2104
neutral	0.67	0.54	0.60	1787
sadness	0.75	0.39	0.51	379
surprise	0.55	0.48	0.51	677
micro avg	0.70	0.61	0.65	5894
macro avg	0.66	0.51	0.57	5894
weighted avg	0.70	0.61	0.64	5894
samples avg	0.65	0.63	0.63	5894

Figure 3: Classification Report of DistilBERT using Ekman Wheel

As a result, the DistilBERT model exhibited a substantial increase in the F1 score, jumping to 0.57. This improvement not only demonstrates the potential of model optimization through strategic data categorization but also aligns with foundational psychological theories on emotional expression, providing a more intuitive and theoretically grounded framework for emotion classification in NLP.

The figure 3 shows the Performance Metrics of 7 broder emotions produced by the DistilBERT model. All the metrics jumped a substantial percentage upon using the Ekman Wheel of Emotion of categorize emotions. We can look at the con-

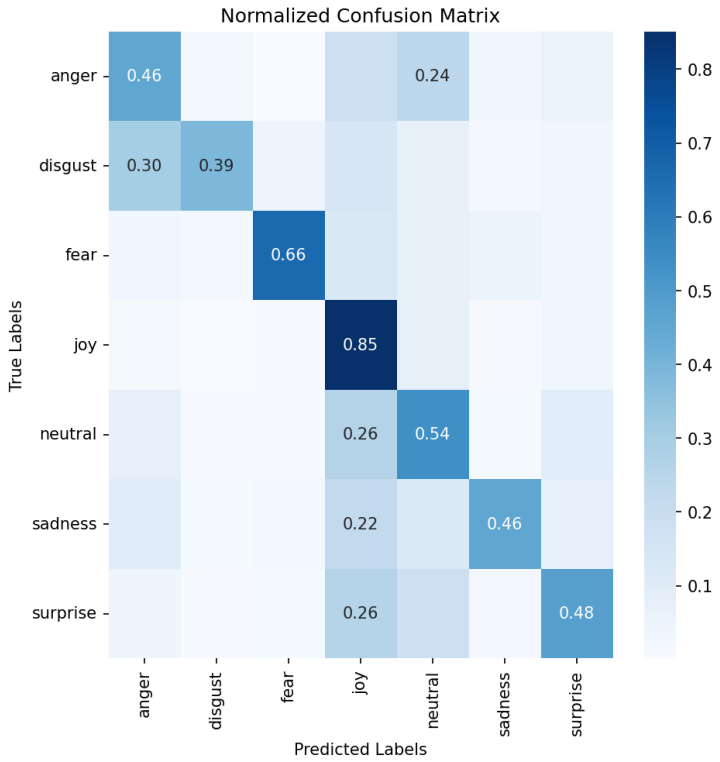

Figure 2: Confusion Matrix of DistilBERT After Using Ekman Wheel to Group Emotions

fusion matrix as well in figure 3 to analyze that the model exhibits a high classification accuracy for 'joy' with a value of 0.85, suggesting a strong ability to distinguish joyful content. fConversely, the model appears to struggle with distinguishing between 'anger' and 'disgust,' as well as 'joy' and 'surprise' states, evidenced by lower true positive rates of 0.30 for anger and 0.26 for neutral, and significant off-diagonal values. These off-diagonal values indicate a considerable degree of confusion between these emotional states, potentially due to overlapping linguistic markers or insufficiently distinctive training data.

6 Further Work

This study only explored transformer models. The one big extension to the study is to incorporate sequence models, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs). These models are particularly adept at handling sequences of data, making them potentially well-suited for the analysis of textual emotional expression. Additionally, exploring different attention mechanisms, beyond the standard attention used in transformer models, could provide deeper insights. This exploration would not only broaden the comparative analysis of model performances but could also uncover innovative approaches to improve emotion classification accuracy.

References

[1] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 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, 2020

[2] Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar Zaïane. Seq2Emo: A Sequence to Multi-Label Emotion Classification Model. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4717–4724, 2021.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019 Bert: Pre-training of deep bidirectional transformers for language understanding. *In 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*

[4] Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the Limits of Text-based Emotion Detection.. *In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2560–2583, Punta Cana, Dominican Republic. Association for Computational Linguistics.*

[5] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. 2020 DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

[6] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. 2019 XLNet: Generalized Autoregressive Pretraining for Language Understanding

Appendix

	precision	recall	f1-score	support
admiration	0.73	0.60	0.66	504
amusement	0.83	0.76	0.80	264
anger	0.69	0.24	0.36	198
annoyance	0.54	0.12	0.19	320
approval	0.74	0.15	0.25	351
caring	0.44	0.40	0.42	135
confusion	0.59	0.17	0.26	153
curiosity	0.48	0.43	0.45	284
desire	0.68	0.25	0.37	83
disappointment	0.90	0.06	0.11	151
disapproval	0.62	0.19	0.29	267
disgust	0.85	0.23	0.36	123
embarrassment	0.53	0.27	0.36	37
excitement	0.54	0.33	0.41	103
fear	0.76	0.54	0.63	78
gratitude	0.94	0.90	0.92	352
grief	0.00	0.00	0.00	6
joy	0.81	0.50	0.62	161
love	0.78	0.86	0.82	238
nervousness	0.43	0.13	0.20	23
optimism	0.70	0.38	0.49	186
pride	0.50	0.06	0.11	16
realization	0.54	0.13	0.21	145
relief	0.67	0.18	0.29	11
remorse	0.57	0.75	0.65	56
sadness	0.80	0.40	0.54	156
surprise	0.70	0.31	0.43	141
neutral	0.72	0.53	0.61	1787
micro avg	0.72	0.45	0.55	6329
macro avg	0.65	0.35	0.42	6329
weighted avg	0.71	0.45	0.52	6329
samples avg	0.50	0.47	0.48	6329

Figure 4: Classification Report of RoBERTa Model

	precision	recall	f1-score	support
admiration	0.77	0.52	0.62	504
amusement	0.80	0.84	0.82	264
anger	0.67	0.33	0.44	198
annoyance	0.76	0.09	0.16	320
approval	0.67	0.18	0.29	351
caring	0.56	0.16	0.25	135
confusion	0.62	0.16	0.25	153
curiosity	0.68	0.05	0.09	284
desire	0.50	0.47	0.48	83
disappointment	0.57	0.11	0.18	151
disapproval	0.45	0.25	0.32	267
disgust	0.92	0.19	0.31	123
embarrassment	0.52	0.41	0.45	37
excitement	0.76	0.24	0.37	103
fear	0.79	0.49	0.60	78
gratitude	0.97	0.87	0.92	352
grief	0.00	0.00	0.00	6
joy	0.61	0.60	0.61	161
love	0.75	0.81	0.78	238
nervousness	0.00	0.00	0.00	23
optimism	0.55	0.60	0.57	186
pride	0.86	0.38	0.52	16
realization	0.89	0.06	0.10	145
relief	0.00	0.00	0.00	11
remorse	0.62	0.55	0.58	56
sadness	0.74	0.44	0.55	156
surprise	0.70	0.40	0.51	141
neutral	0.73	0.54	0.62	1787
micro avg	0.72	0.44	0.54	6329
macro avg	0.62	0.35	0.41	6329
weighted avg	0.71	0.44	0.50	6329
samples avg	0.49	0.46	0.47	6329

Figure 5: Classification Report of XLNET Model

	precision	recall	f1-score	support
admiration	0.73	0.60	0.66	504
amusement	0.76	0.82	0.79	264
anger	0.58	0.42	0.49	198
annoyance	0.45	0.20	0.28	320
approval	0.51	0.28	0.36	351
caring	0.48	0.23	0.31	135
confusion	0.40	0.33	0.36	153
curiosity	0.50	0.34	0.40	284
desire	0.61	0.36	0.45	83
disappointment	0.35	0.21	0.26	151
disapproval	0.38	0.31	0.34	267
disgust	0.68	0.36	0.47	123
embarrassment	0.60	0.24	0.35	37
excitement	0.51	0.35	0.42	103
fear	0.72	0.64	0.68	78
gratitude	0.95	0.89	0.92	352
grief	0.00	0.00	0.00	6
joy	0.67	0.50	0.57	161
love	0.75	0.85	0.79	238
nervousness	0.41	0.30	0.35	23
optimism	0.72	0.45	0.55	186
pride	0.80	0.25	0.38	16
realization	0.48	0.16	0.24	145
relief	0.38	0.27	0.32	11
remorse	0.58	0.66	0.62	56
sadness	0.72	0.42	0.53	156
surprise	0.58	0.45	0.51	141
neutral	0.67	0.57	0.61	1787
micro avg	0.65	0.49	0.56	6329
macro avg	0.57	0.41	0.46	6329
weighted avg	0.63	0.49	0.54	6329
samples avg	0.54	0.52	0.52	6329

Figure 6: Classification Report of DistilBERT Model

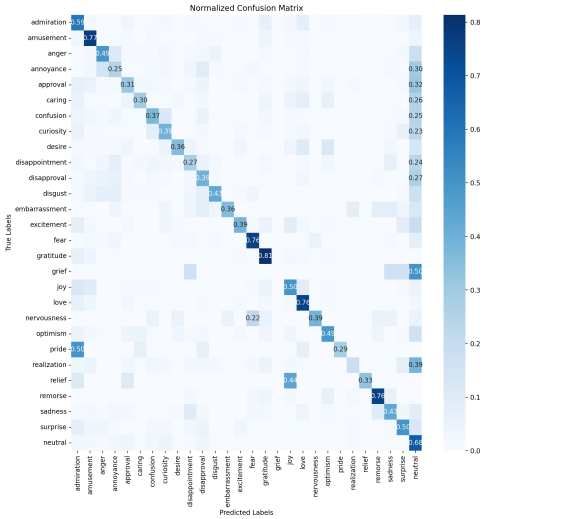

Figure 7: Classification Report of DistilBERT Model