Emotion Detection using BERT-base Model

Alyssa Z, Neha S, Farhan S, Colin S, Ryan N
University of North Carolina at Chapel Hill

Abstract

Text based emotion detection (TBED) has progressed from lexicon and feature engineered models to deep neural architectures on transformers. In this research we analyze standard transformer models for fine grained emotion detected and extend the best forming model into a simple generative user facing setting where arbitrary user inputs are mapped to emotion labels. We fine tuned BERT, RoBERTa, and DistiliBERT on a public Twitter emotion dataset from Kaggle containing 40000 short texts annotated with 13 emotion categories. We first obtain benchmark results in order to compare architectures and establish a baseline. We then retrain the strongest model on the full dataset and deploy it in a generative interface that accepts free form user input and returns predicted emotion labels.

1 Intro/Motivation

Our goal in this project is to compare three widely used transformer models BERT, RoBERTa, and DistilBERT on fine grained emotion dataset and then extend the strongest model into a generative interface. We first compare these models under the same controlled training setup to understand their relative performance in terms of accuracy, loss, efficiency, and training cost. This benchmarking gives insight into how well traditional transformer classifiers perform on short informal text typical of social media.

Because emotion detection and reproduction remain areas where modern AI systems struggle, our motivation is to better understand how current transformer models can extend emotional nuance in real world text. Even the best language models often misinterpret subtle emotional cues, fail to differentiate between closely related emotion categories or collapse rare emotions into more common ones. This makes fine grained emotion detection worth addressing. The dataset used in this study contains 40000 publicly available tweets each labeled by human annotators with one of 13 distinct emotions. By comparing BERT, RoBERTa, and DistilBERT on this large but noisy dataset we aim to identify which model is best suited for applications requiring accurate real time emotion understanding. However, due to the nature of the data - informal Tweets with a wide variety of spelling, slang, and other language use - and the relative size of the dataset considering the quantity of data that is used for popular modern models, we posit that our results may be lackluster.

By training all three transformer models under identical preprocessing steps, hyperparameter settings, and train test splits, we aim to determine which model performs best in terms of accuracy, efficiency, and robustness. This comparison provides insight into how these models handle fine grained emotion classification and helps identify which architecture is best suited for future work in emotion aware NLP applications.

2 Related work

Text based emotion detection is a subfield of sentiment analysis that focuses on identifying specific emotional states expressed in written language. Early research in this area relied on lexicon based and rule based methods such as WordNet Affect, SentiWordNet, and LIWC. These approaches

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

were intuitive because they mapped words to predefined emotion categories, but they struggled with negation, figurative language, and sarcasm. Rule based systems attempted to compensate by incorporating syntactic features but they required large sets of handcrafted rules that did not generalize well to new text.

The rise of machine learning introduced statistical models such as support vector machines, naive bayes, and random forests. These methods used manually engineered features including Bag of Words, TF IDF vectors, and n grams. Although they improved performance and were used successfully in early shared tasks like SemEval, they were limited by their dependence on surface level features and their inability to capture deeper semantics or long range contextual information.

Deep learning methods advanced the field further. Convolutional neural networks captured local emotion indicative patterns, while recurrent neural networks such as LSTMs and GRUs modeled sequential structure which allowed for better interpretation of negation, word order, and evolving emotional cues. Hybrid architectures that combined CNNs and bidirectional GRUs offered more gains. Models like DeepMoji demonstrated the power of large-scale pretraining using emoji prediction as a proxy for emotion. However, RNNs struggled with very long range dependencies and required substantial labeled data.

A major shift occurred with the introduction of the Transformer architecture which uses self attention to capture dependencies more effectively than RNNs. Pretrained transformer models such as BERT, RoBERTa, GPT, and T5 achieved better performance on many NLP benchmarks after fine tuning on small labeled datasets. These architectures quickly became standard for emotion classification tasks.

BERT is a bidirectional encoder that processes all tokens simultaneously which allows it to capture context rich representations. It is trained using masked language modeling and next sentence prediction. RoBERTa builds on BERT by removing the next sentence prediction objective training with ten times more data including Reddit discussions, blogs, news, and CommonCrawl and using dynamic masking so that masked tokens change at every epoch. Dynamic masking forces the model to learn deeper contextual dependencies and prevents memorization of fixed masked positions. These improvements make RoBERTa particularly strong on tasks involving informal emotional language, short text, and subtle emotional cues such as those found in social media.

Recent comparative studies have continued to evaluate transformer based architectures on emotion detection. For example, work on fine grained emotion detection in short text has shown that BERT, RoBERTa, XLNet, ELECTRA, and XLM R can all achieve high accuracy when hyperparameters are tuned carefully. Shared experimental setups in these studies typically include learning rates around 5e 5, batch size 8, dropout 0.1, and train test splits around 70 to 30. Results from these comparisons show that BERT often performs best when trained for approximately nine epochs, although RoBERTa performs competitively and sometimes excels due to its larger training corpus and dynamic masking.

3 Method & set up

We used Tweet Emotion, a publicly available dataset through Kaggle, to implement all three models. The dataset contains 40,000 samples with each row representing a tweet comment and three columns, including: 'Content', which contains the comment from the user, 'ID', the index used to uniquely identify each user, and 'Sentiment', that contains 13 distinct emotion labels.

We preprocessed the dataset by first taking out the Twitter user names formatted with "@...". Next, we constructed a new column to map the emotion label with categorical values, as seen below. A preview of the different labels can be seen below:

Three models were trained for this paper. BERT(Bidirectional Encoder Representations from Transformers), a task-specific model that takes embedded word vectors and aims to predict masked tokens as well as next sentence from previous information. DistillBERT, a compressed version of BERT that uses knowledge distillation to achieve approximately 97% of BERT's performance while being 60% faster and 40% smaller. Lastly, RoBERTa(Robustly Optimized BERT Pretraining Approach), improved from BERT and removed the next sentence prediction task using dynamic masking along with a larger training corpus to learn more contextual representations. All three models modified architecture from Frye et al and used the same parameters: learning rate of 5e-5, batch size 8, seed 21, number of epochs 2, and dataset size 20,000 by randomly selecting half of the dataset and further split into 80% training and 20% testing due to hardware limit.

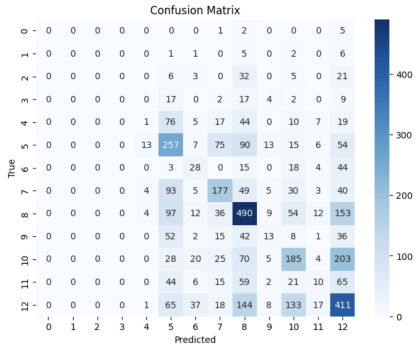
	sentiment_label	sentiment
0	0	anger
1	1	boredom
2	2	empty
3	3	enthusiasm
4	4	fun
5	5	happiness
6	6	hate
7	7	love
8	8	neutral
9	9	relief
10	10	sadness
11	11	surprise
12	12	worry

4 Results & Discussion

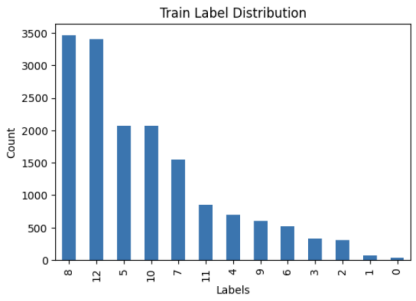
We evaluated all three transformer models using four standard performance metrics: F1-score, recall, precision, and accuracy. Overall, the results indicated modest performance across models with BERT achieving the best results in accuracy but only by a small 1% margin.

Model	F1-Score	Recall	Precision	Accuracy
BERT	0.36	0.39	0.35	0.39
DistillBERT	0.35	0.37	0.35	0.37
RoBERTa	0.32	0.38	0.31	0.38

To better understand the models' behavior, we examined a confusion matrix for each model. Across all three models, the highest performing emotion categories were happiness (label 5), neutral(label 8), and worry(label 12). These classes have comparatively larger representation in the data set which likely contributes to stronger performance. In contrast, labels 0 - 4 like anger, boredom, empty, enthusiasm, and fun received almost no correct predictions. Many of these categories had very small sample sizes and the tail heavy imbalance caused the models to underpredict them, or fail to predict them entirely.



To research this further, we analyzed the label distribution of the 20000 sample subset used for training. The distribution is highly imbalanced with several emotions appearing only a few dozen times while others appear hundreds of times. This imbalance combined with the limited number of training epochs and reduced dataset size due to hardware constraints likely explains the poor performance on the minority categories.



Lastly, we implemented a generative model interface that accepts user input and returns a predicted emotion label. Because the models performed poorly on the test set, we did not expect strong results in this interactive setting and the generative outputs reflected this. The models predicted happiness, neutral, or worry which were the same labels that dominated both the training distribution and the confusion matrix while failing to correctly assign less frequent emotion categories. This reinforces our earlier observation that data imbalance and limited training constrained the model performance. To improve results, the dataset would need normalization or rebalancing so minority categories are more represented. Also, increasing the number of epochs beyond 2 while computationally expensive would likely allow the models to learn better representations. If we were to implement these two improvements, we would expect both the test set performance and the generative predictions to improve significantly.

5 Conclusion

In this study, we conducted a comparative evaluation of BERT, DistilBERT, and RoBERTa on an emotion classification task to assess their effectiveness under constrained training conditions and created a generative interface that accepts free form user input and returns predicted emotion labels. While BERT demonstrates slightly stronger accuracy performance and DistiliBERT is the fastest among all three models, all the models show clear difficulties in learning fine grained emotional distinctions under limited data and class imbalance (see graph “Train Label Distribution”). These findings suggest that improving performance would either require additional training epochs, a larger portion of the dataset, or rebalancing to fix oversampling or weighted loss.