Performance evaluation of Reddit Comments using ML and NLP methods in Sentiment Analysis

Xiuyuan Qi
Shanghaitech University
qixy1@shanghaitech.edu.cn

Zixin Teng
Shanghaitech University
tengzx@shanghaitech.edu.cn

Xiaoxia Zhang
Shanghaitech University
zhangxx5@shanghaitech.edu.cn

Abstract

Sentiment analysis, crucial in machine learning, particularly for platforms like Reddit, suffers from small-scale, coarse emotion datasets. We urgently need expansive, fine-grained datasets, so we utilize the GoEmotions dataset for our study, evaluating various methods on 58,000 comments. This includes traditional classifiers (Bayes, SVM) and recent models like BERT, RoBERTa and GPT based model. Our analysis concludes that the RoBERTa model outperformed the baseline significantly. Our experiment result highlights the substantial potential and importance of the RoBERTa model in fine-grained sentiment classification.

1 Introduction

Sentiment analysis represents a pivotal domain within research, particularly within commercial contexts, owing to its significance in delineating user profiles on social platforms. Despite the considerable contributions of Natural Language Processing (NLP) researchers in developing sentiment classification datasets spanning various domains such as news and Twitter, these datasets frequently present limitations in both scale and granularity. Primarily, they adhere to binary or Ekman's emotion classification schemes. The binary taxonomy oversimplifies emotional nuances by categorizing sentiments solely into positive and negative types, thus lacking precision and accuracy. In the framework of discrete emotion theory, it is posited that all humans possess an intrinsic set of basic emotions that are universally recognizable across cultures. The seminal work of Paul Ekman and his colleagues in their 1992 cross-cultural study delineated six fundamental emotions: anger, disgust, fear, happiness, sadness, and surprise, forming the classical Ekman's classification standard. However, the delineation of emotions into six categories fails to adequately address the intricate nature of contemporary emotion classification tasks, particularly within the dynamic landscape of social media where a vast array of emotions is expressed, often defying easy categorization. Consequently, there arises an urgent necessity for the development of a comprehensive, large-scale dataset capable of nuanced emotion classification, serving as the foundational premise of our study.

In addressing these challenges, we conducted a comprehensive evaluation of multiple datasets, ultimately selecting the GoEmotions dataset developed by Google as the foundational corpus for our study. Comprising 58,000 Reddit platform comments meticulously annotated to encompass 27

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

emotion categories or "neutral" sentiments, GoEmotions caters to the diverse array of information typified by Reddit as a comprehensive, large-scale social news forum. Diverging from Ekman's emotion classification taxonomy, GoEmotions incorporates a plethora of positive, negative, and ambiguous emotion categories, rendering it particularly suitable for downstream conversational understanding tasks necessitating nuanced emotion comprehension. Besides, the Reddit platform has hitherto not garnered significant attention from researchers in the field of sentiment recognition and prediction. The GoEmotions dataset, however, stands as the inaugural dataset specifically tailored to this platform, thereby stimulating our group members' interest in employing a wider array of machine learning models to train and evaluate classification results. They also train a bidirectional LSTM as an additional baseline but performs significantly worse than BERT.

[Demszky et al., 2020] presented detailed work for building the GoEmotions dataset and showing how to demonstrate the high quality of the annotations via Principal Preserved Component Analysis. They conduct transfer learning experiments with existing emotion benchmarks to prove that GoEmotions generalizes well to other domains and different emotion taxonomies. Furthermore, Google team tried to use a fine-tuned BERT based model originating from [Devlin et al., 2019] to test the dataset and achieved an average F1-score of .46 across the proposed taxonomy of 27 emotion categories, leaving much room for improvement.

Based on the research conducted by the Google team and the comprehensive preprocessing efforts applied to this dataset, we intend to utilize the results obtained from this dataset in conjunction with the Google BERT model as the baseline for our model evaluation. Additionally, we aim to initially implement and evaluate the performance of various Bayesian models and Support Vector Machines (SVMs), among other machine learning models covered in the curriculum. Furthermore, we will focus on large language models based on the RoBERTa model and the GPT interface. Our objective is to iteratively fine-tune and assess the classification accuracy, computational efficiency, and other relevant metrics of these models, with the ultimate aim of achieving superior training outcomes compared to the original Google models.

2 Methodology

2.1 Machine Learning Method

Our evaluation process commences with the utilization of the GoEmotions dataset. We aim to train a variety of machine learning models covered in our curriculum, alongside advanced natural language processing models. These models encompass a comprehensive pipeline comprising tokenization, embedding, transformation, and post-processing steps, such as applying sigmoid activation functions. Subsequently, all models undergo one-hot encoding to generate a series of outcomes, which are evaluated using diverse performance metrics. The specific workflow of this process is depicted in the diagram below.

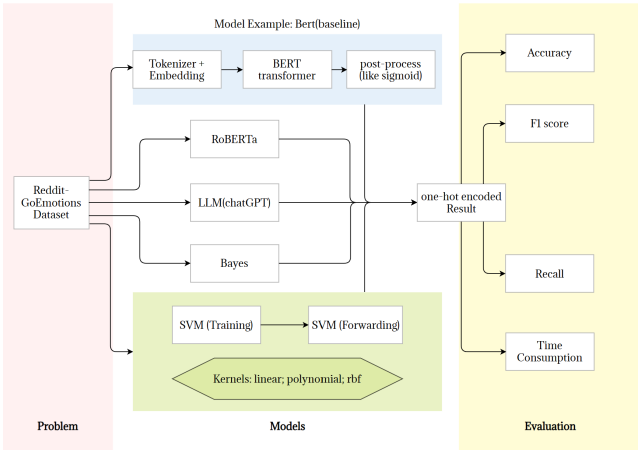

Figure 1: Processing Methods

Bayes+MLE/KNN The Bayesian model, based on Bayesian probability theory, provides a probabilistic framework for sentiment classification tasks. It calculates sentiment label probabilities using Bayesian inference, enabling principled decision-making. By addressing uncertainty, it offers interpretability and resilience, valuable for text data analysis. Integrating Bayesian modeling with Maximum Likelihood Estimation (MLE) and K-Nearest Neighbors (KNN) methods enhances our approach, particularly in the Complement Naive Bayes Classifier.

SVM In sentiment classification prediction, Support Vector Machine (SVM) is a prominent supervised learning method widely used in text analysis tasks. SVM maps data points into high-dimensional spaces to find an optimal hyperplane for classification. Its role involves effectively discerning sentiment orientations in text, handling high-dimensional feature spaces, and exhibiting resilience against noise. By utilizing different kernel functions like linear, polynomial, and radial basis function (RBF), SVM adapts to diverse feature mappings for tailored sentiment analysis, enhancing classification accuracy. We will test all the above three kernels.

2.2 Natural Language Processing Method

BERT The BERT (Bidirectional Encoder Representations from Transformers) model serves as our baseline, representing a focal point of research by the Google team. They enhanced the model by incorporating a dense output layer to facilitate multi-label classification, accompanied by the utilization of a sigmoid cross-entropy loss function. Google emphasized the importance of training for a minimum of 4 epochs to effectively capture the underlying data patterns, while cautioning against the risk of overfitting with prolonged training durations. Leveraging the fine-tuned hyperparameters established by [Devlin et al., 2019] alongside the adjustments recommended by Google for the batch size set at 16 and a learning rate of 5e-5, is anticipated to yield optimal performance in the GoEmotions dataset.

RoBERTa The RoBERTa model, proposed in [Liu et al., 2019], is an extension of Google's BERT model introduced in 2018. Building upon BERT, RoBERTa modifies crucial hyperparameters, such as removing the next-sentence pretraining objective and employing larger mini-batches and learning rates during training. This refinement aims to enhance the model's robustness and performance. In sentiment classification tasks, RoBERTa's enhanced pretraining strategies and robust architecture may offer improved capabilities for accurately capturing nuanced sentiment expressions and achieving state-of-the-art performance in sentiment classification benchmarks.

LLM(GPT) The GPT (Generative Pre-trained Transformer) model, including GPT-3.5 Turbo from OpenAI, is a cutting-edge language generation model based on transformer architecture. Unlike BERT, which focuses on bidirectional contextual understanding, GPT generates text sequentially, predicting the next token based on preceding ones. Although not originally designed for sentiment classification, GPT can be adapted for such tasks by conditioning on sentiment-related prompts. In sentiment analysis, GPT generates text expressing sentiment, which can then be categorized. We'll implement a framework using GPT-3.5 for text sentiment analysis and evaluate its performance.

3 Experiment

3.1 Dataset

As we mentioned in introduction, the GoEmotions dataset is the largest manually annotated dataset of 58k English Reddit comments, labeled for 27 emotion categories or Neutral. The training dataset employed for model training comprises instances where a consensus exists among at least 2 raters. These datasets have no header row and are structured with three columns: "text," representing the textual content; a comma-separated list denoting emotion IDs, indexed from 1 to 28; and an identifier for each comment.

Specifically, the emotion categories are: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise. They are indexed by integers 1 to 27. And the number 28 stands for the nuetral type.

The GoEmotions dataset offers several advantages. Firstly, a significant majority of the examples (83%) possess a singular emotion label, ensuring clarity in annotation. Additionally, a high level of agreement among raters (94%) on these single labels enhances the dataset's reliability. To further enhance data quality, emotion labels selected by only one annotator are filtered out, resulting in the retention of 93% of the original data. Finally, the dataset is thoughtfully partitioned into train (80%, 43410 samples), dev (10%, 5426 samples), and test (10%, 5427 samples) sets, facilitating robust model evaluation. In addition to Google's 28 finer-grained sentiment categories (referred to as 'the original task'), we've kept Ekman's six-category classification and grouped sentiment into positive/negative/neutral/ambiguous categories. This enables evaluating models' abilities across different levels of classification precision.

3.2 Result

Evaluation Metrics We will use four metrics to test our models: Accuracy, F1 Score, Recall and Time Consumption. The accuracy, in the context of classification tasks, refers to the proportion of correctly classified instances out of the total number of instances evaluated.

$$\text{Accuracy} = \frac{\textit{Number of Samples with Correct Classification}}{\textit{Total Number of Samples}}$$

Precision measures the proportion of true positive predictions among all positive predictions made by the model. And recall measures the proportion of true positive predictions among all actual positive instances in the dataset. The definition of F1 Score is:

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

As for the time consumption, we record the rounded average time it takes for different models to complete multiple equal classification tasks.

Table 1: Result of different models for original task

Method	Model	Accuracy	F1 Score	Recall	Time(s)
Machine Learning	Bayes+MLE	0.023	0.01	0.03	7
	Bayes+KNN	0.009	0.01	0.02	79
	SVM (linear)	0.010	0.01	0.01	5
	SVM (polynomial)	0.016	0.03	0.02	6
	SVM (rbf)	0.015	0.01	0.02	6
Natural Language Processing	BERT	0.42	0.57	0.61	408
	RoBERTa	0.45	0.60	0.64	421
	LLM(GPT)	0.02	0.013	0.05	>2000

Analysis Based on the experimental results, machine learning models exhibit significant disparities from the strong baseline established by Google's experiments. This outcome is unsurprising as we anticipated these algorithms lack specialized adaptability for sentiment classification tasks. These models show a capability to narrow the gap with the baseline in tasks with fewer labels. The longer processing time observed in natural language processing models is attributed to their complexity. RoBERTa model demonstrates the most outstanding performance overall, effectively accomplishing classification tasks across more than twenty labels, thus highlighting its optimized characteristics built upon the BERT model foundation.

4 Conclusion

In this study, we focused on the GoEmotions dataset to evaluate sentiment classification tasks on Reddit platform comments using various machine learning and natural language processing models. RoBERTa emerged as the top performer, thanks to its strengths in contextual understanding, extensive pretraining, and fine-tuning capabilities. Future research will delve into disintegration experiments on models like RoBERTa to better understand their principles. Additionally, the advent of more complex fusion models may reshape sentiment classification by leveraging the strengths of multiple advanced models.

References

[1] Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade and Sujith Ravi. "GoEmotions: A Dataset of Fine-Grained Emotions." Annual Meeting of the Association for Computational Linguistics (2020).

[2] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz and Jamie Brew. "HuggingFace's Transformers: State-of-the-art Natural Language Processing." ArXiv abs/1910.03771 (2019): n. pag.

[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] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692