Emotion Recognition Uncertainty Modeling of Internet Comments

Yana Zlatanova
yana.zlatanova.gold@gmail.com

André Plancha
andre.plancha@hotmail.com

Vérane Flaujac
vflaujac@gmail.com

Abstract

Understanding user sentiment and emotion is a growing challenge in machine learning, especially for enabling chatbots to respond with empathy and relevance. In this study, we focus on classifying emotions in short sentences using the GoEmotions dataset, which contains Reddit comments labeled with 28 distinct emotions. Throughout our analysis, we discovered that emotion classification is inherently subjective. Multiple annotators often assigned different labels to the same comment, and some emotions appeared ambiguous. After preprocessing the data and performing exploratory visualizations to understand its structure, we trained a model using distillBERT for emotion classification. We then evaluated the model's performance using various metrics to assess its effectiveness. **Our work also explores key questions such as: Which emotions are represented? Does the dataset show demographic bias? And is the model accurate enough for deployment in real-world chatbot systems? This study builds upon existing research while offering our own perspective on implementing and analyzing emotion recognition models.**

Keywords

Emotion Recognition, NLP, GoEmotions, Internet Comments, Machine Learning, HuggingFace, DistillBERT, Evaluation Metrics

1 Introduction

Emotions play a vital role in human communication as the flavor of speech that determines the meaning beyond the literal translation. This makes emotion recognition an important subdomain of natural language processing (NLP), as it enables machines to better interpret human text and speech which is particularly useful in mental health monitoring, customer feedback analysis, and chatbots [16]. However, emotion recognition comes with its challenges. Traditional machine learning models require extensive feature engineering, making them inefficient and unable to keep up with the complexity of human language. On the other hand, although deep neural networks offer better performance, they typically require large amounts of labeled training data which is difficult to obtain [2].

One solution this challenge is transfer learning, which offers pre-trained transformer-based models such as BERT, RoBERTa, DistilBERT, and XLNet. These models have been trained on vast amount of data and are capable of understanding complex linguistic structures and contextual relationships. Their knowledge can be transferred by fine-tuning them on a specialized emotion-labeled dataset to achieve efficient emotion classification without training from scratch [2].

In this study explored emotion classification in text using DistilBERT-based models [18], a smaller and faster variant of BERT, to answer the following research questions:

- What are the challenges in emotion recognition with a human annotated dataset?
- Can fine-tuned models reliably classify emotions from Reddit comments, despite annotator disagreement?
- How does emotion classification performance differ across taxonomies?
- What is the effect of different loss functions on the fine-tuning performance of DistilBERT?

These models were fine-tuned on the GoEmotions dataset [7], which contains approximately 58,000 Reddit comments annotated with 28 emotion labels.

This report serves as a comprehensive overview of our study, detailing the dataset and our exploration of it, previous research done on both the GoEmotions dataset and in sentiment and emotion analysis the transformations applied and task formulation, the model architectures and performance metrics, and the results of our experiments. In short, it describes the various issues we found with the dataset, how we solved them, including how we decided to incorporate human annotator disagreements, how we decided to model and evaluate the models, and the general conclusions of the study.

2 Dataset

GoEmotions is the largest human-annotated emotion dataset, with multiple labels per comment to ensure quality [7]. This section outlines how the dataset was collected, annotated, and processed, following the original scientific publication by its creators [7].

A key innovation is its 27-emotion taxonomy, illustrated in Figure 2, based on modern psychological research and going beyond Ekman's six basic emotions. The dataset includes English-only Reddit comments from subreddits containing at least 10k comments.

Sample Text	Label(s)
OMG, yep!!! That is the final answer. Thank you so much!	gratitude, approval
I'm not even sure what it is, why do people hate it	confusion
Guilty of doing this tbph	remorse
This caught me off guard for real. I'm actually off my bed laughing	surprise, amusement
I tried to send this to a friend but [NAME] knocked it away.	disappointment

Figure 1: Snippet of the Goemotions Dataset.

2.1 Annotation Process

To ensure annotation quality, each comment was reviewed by multiple raters [7] who categorized the comment into to emotions. Initially, three annotators assessed each comment. If there was no agreement on at least one emotion label, two additional annotators were assigned. All raters were native English speakers from India and they were presented with the comments without author or subreddit information. [1]

2.2 Taxonomy

It is important for us to know how was the data set collected, put together and cleaned for us to be able to interpret the results correctly.

The 27-category emotion taxonomy of the dataset was inspired by modern psychological research which is far beyond the traditional six basic emotions — joy, anger, fear, sadness, disgust, and surprise — originally proposed by Ekman [7].

Comments containing offensive or adult language were removed, except for vulgar comments, which were kept to help study negative emotions. Comments with offensive content toward minorities were manually removed. Only comments with 3 to 30 tokens (including punctuation) were retained. Various techniques were applied to balance the dataset and reduce emotion overrepresentation. Additionally, personal names and religion terms were masked with [NAME] and [RELIGION] tokens, respectively. Note that raters saw the original, unmasked comments during annotation.

3 Literature review

The GoEmotions dataset was introduced by Demszky, D. et al. [7], which outlines the motivation, processes, and tools used to create the dataset we are using, along with experiments showcasing its effectiveness. The GoEmotions dataset was introduced to address the lack of sufficiently large datasets for language-based emotion classification and the limitations of existing emotion taxonomies, which typically use limited emotion taxonomies, such as Ekman's 6 emotions [8]. Demszky, D. et al. claims they created the largest human-annotated dataset of 58k carefully selected Reddit comments, labeled with **27 emotion categories or Neutral**, as shown on Figure 2, drawn from popular English subreddits. The dataset stands out for its richer taxonomy, which includes a more diverse range of positive, negative, and ambiguous emotions; in contrast, unlike Ekman's taxonomy includes only one positive emotion (joy).

The paper explains how the dataset was constructed and presented a baseline BERT-based model for emotion prediction, achieving a F_1 of 0.46 over the proposed 27 emotions taxonomy, but performed better with a 0.64 score using an Ekman-style grouping into six emotion categories and 0.69 using a simple sentiment grouping (positive, neutral, negative) [7]. These results suggest that the broader the emotion group, the better the accuracy. This new taxonomy proposal inspired us to explore different emotion categories and also confirmed that a BERT based model would be suitable for our aims. The Dataset has been used and analyzed in following studies, with Wang, K. et al. [20] achieving comparable or better results, while comparing different models and a fine-tuned BERT model. For this study, we decided to use DistilBERT [18] as our BERT based model, as DistilBERT is a streamlined version of BERT developed by a Hugging Face team that achieves significant reductions in model size and inference time of BERT while maintaining most of its performance.

The paper also examines limitations of the dataset and ways to address them, such as the big class imbalance and biases present in the dataset. We intend to to expand on this notion in Section 2. To help with class imbalance, Wang, K. et al. [20] explores data augmentation methods such as Easy Data Augmentation, BERT Embeddings, and Bert-based ProtAugment; however, the improvement was marginal, with an increase of 0.027 from the F1 score of the original dataset. Nevertheless, they still achieved a significantly better performance on underrepresented emotion labels, which they attributed to using 10 training epochs instead of Demszky, D. et al.'s [7] 4 epochs.

Emotion Recognition in natural language processing is a complex field, with different nomenclatures, models and frameworks [16]. For this report, we use emotion recognition, emotion prediction and emotion classification interchangeably as the classification of one or multiple emotions portrayed in a specific written text. Following Plaza-del-Arco, F.M. et al.'s [16] study, we also outline that we follow the discrete model of emotions, where each emotion is distinct between each other; however, opposed to other studies in emotion classification [6], we will take advantage of the dataset's collection methodology [7] and use both the multi-label facet and human label variation available to avoid masking the degree to which annotators disagree [6]. We will elaborate this further in the sections that follow.

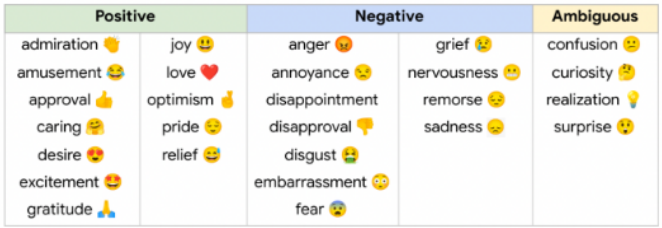

Figure 2: Emotions Comprised in the Dataset.

4 Methodology

In this section, we outline the approach and tools used to carry out our study, from data preprocessing to model training and evaluation. The entire workflow consists of Data exploration, preprocessing from data driven decisions, modeling, and evaluating. Additionally, we developed a user interface to visualize the output of these models. A screenshot of the UI can be visualized in Figure 3. The entire process, including model implementations, is available in https://github.com/yanazlatanova/emotion-recognition.

We used the Hugging Face Transformers library to implement DistilBERT. It is well suited for emotion classification tasks due to its ability to preserve much of BERT's performance while being more efficient. Before feeding the text into the model, we used Hugging Face's tokenizer to convert sentences into token IDs. The tokenizer handles out-of-vocabulary words by breaking them into word units, ensuring even rare or misspelled terms are processed effectively.

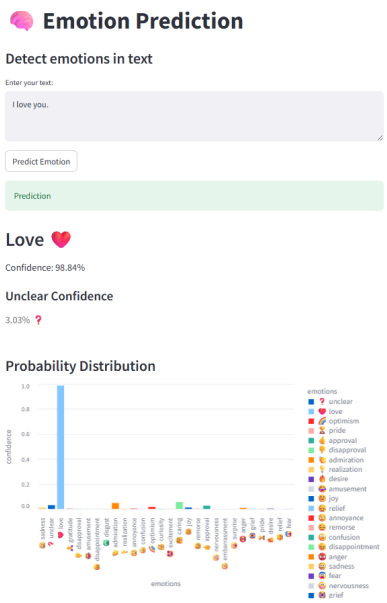

Figure 3: User interface for emotion recognition of input text build using Streamlit.

Training was conducted using PyTorch as the backend. We split the dataset into training, validation, and test sets to ensure a fair assessment of the model's generalization capability. We trained the model and evaluated it using standard classification metrics that gave us information on the performance of the model, the general error and check for possible overconfidence.

4.1 Data Exploration and Pre-processing Decisions

Our initial data exploration, along with a review of the paper by the dataset creators [7], helped us better understand the dataset and informed our pre-processing strategy.

We made a graph of the distribution of the length of the comments as we can see on Figure 4. We found that the distribution was normal and that most of the comment length was around 60 characters. When we selected the comments of length inferior to 4 characters, it showed "yes/no" comments or emoticons eg. such as *:^)* and *XD*.

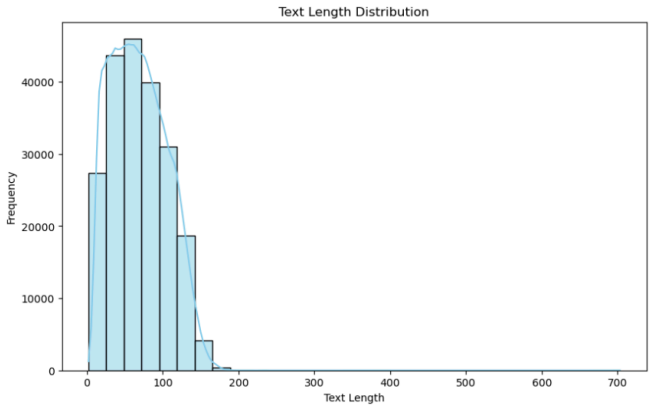

Figure 4: Comment Length Distribution

We found that each comment/text had multiple duplicates, each rated by different annotators. While there were no missing values, some rows were marked as either having no assigned emotion or labeled as "Neutral", often reflecting that the emotion of the comment was unclear. According to Demszky, D. et al. [7], "If raters were not certain about any emotion being expressed, they were asked to select Neutral. We included a checkbox for raters to indicate if an example was particularly difficult to label, in which case they could select no emotions."However, in some cases, raters still assigned an emotion to a comment that another rater found unclear or believed expressed no emotion. This introduced challenges in interpreting such instances.

To further analyze the multiple raters per text, we plotted the number of raters per comment, which can be seen in Figure 5. We could see that most texts were rated by three annotators, but some had only one or two.

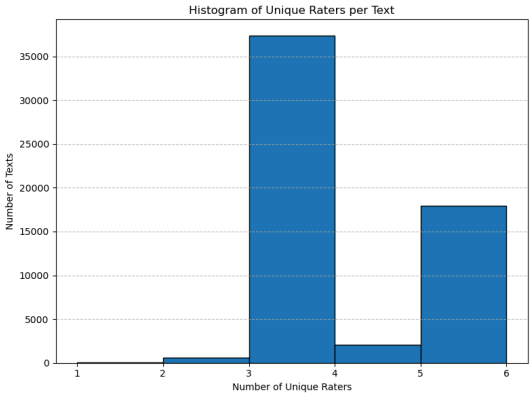

Figure 5: Distribution of the number of unique raters per comment/ text.

Additionally, these raters often disagree in the emotions they assign, even in seemingly contradictory emotions. For example, the comment *"Definitely was a nononoyes[1] for me there lol, I'm a horrible person"* got as fear, amusement, approval and disgust, from 5 different annotators. This noise, while problematic, is expected, since emotion classification is fairly subjective and there's multiple plausible answers [15], annotators' lived experiences influence their interpretations [6], and context is often necessary to access correctly the emotion (which the annotators did not have access to [7]), and domain knowledge (which most annotators might've not had, because of the "ever-evolving ethos" of Reddit's culture, both site-wide and within each subreddit [13]). On top of this, this dataset has been criticized before for its reliability [1, 4], with specific examples from Edwin Chen [1] outlying issues on comments containing profanity, sarcasm, internet style conventions, and culturally specific references. From, this we realize that our model may struggle to accurately predict emotions in future inputs, especially on inputs that contain these.

Based on these facts, we decided to:

- **Label unclear cases consistently**: We treated both "Neutral" and empty emotion labels as "Unclear", united under the *unclear* label, since in both situations raters were unable to confidently identify an emotion.
- **Filter by rater count:** The dataset was filtered to include only comments with at least three raters, ensuring more reliable and confident annotations. As shown in Figure 5, multiple comments had only one or two raters.
- **Aggregated Ratings**: Since identical comments were often assigned different labels by different raters, we chose to

[1]nononoyes is a name of a subreddit dedicated to sharing videos that depict situations initially appearing to go wrong ("no, no, no...") but ultimately result in a surprisingly positive or successful outcome ("yes!").

aggregate the emotion ratings for each unique comment, embracing human label variation [15]. For instance, if the comment "this is adorable" received the following annotations:

- **Rater 1:** [admiration, joy];
- **Rater 2:** [admiration];
- **Rater 3:** [amusement].

The aggregated label distribution would be:

- **admiration:** 0.67 (2 out of 3 raters);
- **amusement:** 0.33 (1 out of 3 raters).
- **joy**: 0.33 (1 out of 3 raters)

This aggregation gives more weight to emotions confirmed by multiple raters while still capturing the presence of less-agreed-upon emotions. It preserves the richness of the label diversity without resorting to semi-supervised learning. Essentially, it gives us a confidence level of the emotions per text. We found this preferable to majority vote aggregation since it preserves different annotators' perspectives [6], while avoiding the different issues that arise with assigning a "ground truth" label to a text [6, 14, 15]

While annotator disagreement and emotion uncertainty has been explored before [4, 6, 9, 21], this way of transforming the dataset seems to be a novel approach in treating annotator disagreement in both the GoEmotions dataset and emotion recognition, as it seems quite more to aggregate label disagreements into a single one [9] since it introduces uncertainty into emotion recognition, transforming our multi label classification task into multi label regression problem, using "soft" labels instead. Arguably, this also gives a hierarchical label structure to the emotions [10], enabling this dataset to be used for point-wise learning to rank tasks. While this is not our priority, our metrics and models will take this into account as well.

Finally, we also replicated the dataset into 3 separate datasets, where the emotion taxonomies previously mentioned were recreated using the same aggregation used by Demszky, D. et al. [7]. These will be reflected in the results.

4.2 Uneven emotion categories

The GoEmotions dataset has an uneven number of comments for each emotion category, as shown in Figure 6, which makes some emotions underrepresented in the dataset, potentially resulting in a worse performance on underrepresented emotions. Because of this unevenness, our performance metrics need to take imbalance into account, to make us undertand the performance of the model in underrepresented emotions, since this is especially important in predicting emotions reliably in the context of single emotion prediction. This is also something Wang, K. et al. [20] noticed while fine-tuning a BERT model on the GoEmotions dataset, where the "grief" label, which has the least sample size in their training set, achieved the worst performance across different evaluation metrics [20].

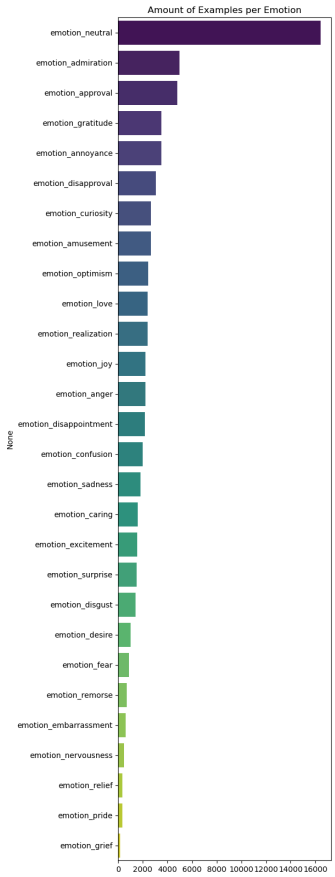

Figure 6: Distribution emotion categories.

4.3 Emotion groups Correlation

We can see in Figure 7 that some emotions are correlated as the dark color tell us. Annoyance and anger are very much linked, as well as nervousness and fear, sadness and disappointment or joy and excitement to cite a few examples. It can be explained by the fact that some emotions are verbally implicit and need more context to be interpreted.

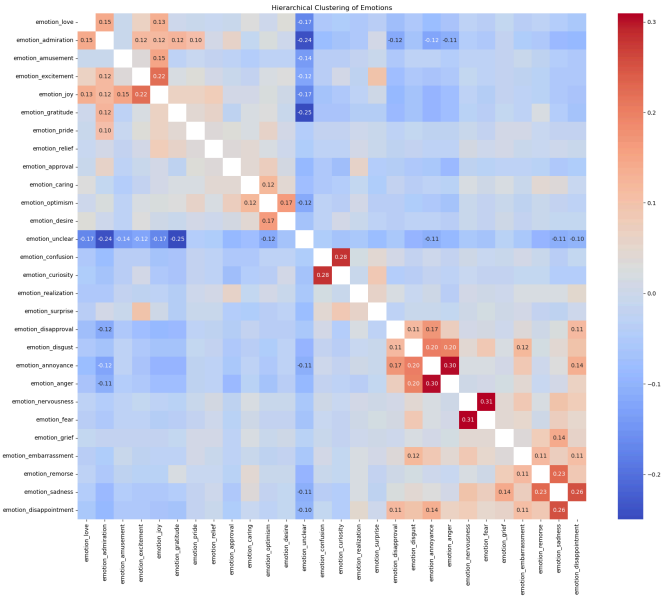

Figure 7: Confusion matrix

In analyzing the GoEmotions dataset, Alba Curry et al. [16] previously employed several techniques to better understand the consistency of emotion labeling. Through hierarchical clustering, they discovered that emotions naturally group by intensity and sentiment polarity, for example "ambiguous" emotions like surprise, cluster closer to positive emotions. To evaluate rater agreement and uncover deeper patterns, they applied Principal Preserved Component Analysis (PPCA), which showed all 27 emotion categories to be highly distinct, which is an unusually strong result in emotion research. To further explore how emotions are organized, they used t-SNE, a dimensionality reduction method, to visualize how emotion labels relate in space. Lastly, they analyzed the linguistic features tied to each emotion by examining which words were statistically most associated with each category. They found that emotions with clear lexical markers—like gratitude being linked to "thanks" showed higher inter-rater agreement, while more context dependent emotions such as grief or nervousness were harder to label consistently. These findings highlight both the richness and the limitations of text-based emotion annotation.

4.4 Emotion Taxonomies/Grouping emotions

Emotions are complex and multifaceted, and researchers have proposed various taxonomies to categorize and study them effectively. One of the most influential models is Paul Ekman's basic emotion theory (Ekman, 1992), which identifies six universal emotions—anger, disgust, fear, happiness, sadness, and surprise—based on cross-cultural facial expression studies. Ekman's framework is widely used in psychology and computational emotion analysis for its simplicity and empirical grounding.

Another widely cited taxonomy is Plutchik's Wheel of Emotions (Plutchik, 1980), which organizes emotions in a circular structure based on intensity and similarity. Plutchik identifies eight primary emotions—joy, trust, fear, surprise, sadness, disgust, anger, and anticipation—each with varying degrees and opposites. This model is particularly useful in visualizing relationships between emotions and understanding how complex emotions arise from combinations of more basic ones.

Beyond these foundational models, more recent work by Bostan and Klinger (2018) aggregates 14 commonly used emotion classification schemes to analyze and unify emotion annotation practices in NLP. Their comparative study emphasizes how emotion categories vary across datasets, revealing discrepancies in granularity, terminology, and theoretical underpinnings. This work underscores the importance of standardizing emotional labels, especially in machine learning contexts, where inconsistent categorization can lead to ambiguous or biased model predictions.

Together, these taxonomies reflect the diversity in how emotions can be defined, labeled, and interpreted—highlighting the challenges and considerations in building accurate emotion recognition systems.

In our study, we chose to adopt Ekman's six basic emotions as the foundation for our classification task. This decision was made to simplify the emotion space while maintaining a strong grounding in psychological theory. Additionally, we introduced a seventh category labeled as "unclear" to account for comments that are either ambiguous, emotionally neutral, or inconsistently labeled by human raters. This category helps manage noise in the dataset, especially given the subjectivity involved in interpreting emotions from short texts.

By narrowing our classification to these seven categories, we aim to strike a balance between theoretical soundness and practical model performance, while acknowledging the limitations of emotional ambiguity in natural language.

4.5 Performance Metrics

Due to the nature of uncertainty prediction and emotion complexity, it is critical that we don't focus on single metrics to evaluate the performance of our models, as that "gives no indication on how reasonable a model is, yet alone how confident and trustworthy it is" [15]. For that end, we implemented multiple different performance metrics, to measure the performance of our models in different aspects.

As a regular regression metric, we used the Mean Binary Cross Entropy (MBCE). This metric is very common in this type of task, and will give us information on the general error of the model, as well as as overconfidence. This is calculated as the following:

$$\begin{aligned} \mathrm{BCE}(y_e, \hat{y_e}) &= -(y_e \log(\hat{y_e}) + (1 - y_e) \log(1 - \hat{y_e})) \\ \mathrm{MBCE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{\sum_{e=1}^{\#\boldsymbol{y}} \mathrm{BCE}(y_e, \hat{y_e})}{\#y} \end{aligned} \tag{1}$$

As regular multi label classification metrics, we're going to use 2 different macro-averaged F_1 metrics, differentiating on the true label definitions we use. As the predicted labels, we decided on using 0.5 as the threshold for a positive or negative prediction; for the ground truth, for the metric F_1^{any}, we say a label is positive if any of the annotators rated as such, and for the metric F_1^{conf}, we say that the emotion with the most confidence, and every emotion with more than 0.8 confidence, are positive. The macro averaged F_1 is a useful classification metric in this case because it takes into account the imbalance of the dataset, giving us more insight on the ability of the model to predict less common emotions.

Additionally, we employ a weighted mean squared error (WMSE), with a weight function designed to penalize errors on higher ground truth confidences. This metric is designed to check the underconfidence of the model, incentivizing the model to be more bold with their predictions, while still taking into account high errors. It can be calculated as the following:

$$\mathrm{WMSE}(y, \hat{y}) = e^{y}(y - \hat{y})^2 \tag{2}$$

This weight function was considered because of its exponential increase, since higher confidence means more annotators agreed on the emotion it seemed appropriate. Table 1 shows other considered weight functions.

Table 1
Different weight functions

$y+1$	e^y	2^y	$e^{\frac{y}{2}}$
1.00	1.00	1.00	1.00
1.20	1.22	1.15	1.11
1.40	1.49	1.32	1.22
1.60	1.82	1.52	1.35
1.80	2.22	1.78	1.49
2.00	2.72	2.00	1.65

Finally, we employ the Normalized Discounted Cumulative Gain (nDCG), which is a common metric in information retrieval and in learning to rank tasks. The metric measures the quality of a ranked list by comparing the actual ranking to the ideal one, rewarding highly relevant items that appear earlier in the list, while still taking ground truth scores/confidences into account. While our task is not strictly a learning to rank task, the metric enables us to understand if the model is correctly giving higher confidence scores to higher confidence emotions even in low confidence scenarios, while still taking into account the ground truth confidence scores. While making sure that $\hat{\boldsymbol{y}}$ is sorted before calculating the DCG, this metric is calculated as the following:

$$\begin{aligned} \mathrm{DCG}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \sum_{i=1}^{\#\hat{\boldsymbol{y}}} \frac{2^{y_i} - 1}{\log(i+1)} \\ \mathrm{nDCG}(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{\mathrm{DCG}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\mathrm{DCG}(\boldsymbol{y}, \boldsymbol{y})} \end{aligned} \tag{3}$$

4.6 Modeling

Our models consist on fine-tuning DistilBERT to a multi-label regression task. Specifically, the text is tokenized (using DistilBert's tokenizer) and then the tokens and attention mask are fed into DistilBert's; the output of that model is fed into 2 dense layers, separated by a dropout layer, and concludes with a sigmoid activation layer over the output logits, to draw predictions per emotion[2]. To train these models, the DistilBert layers are frozen, to not only reduce computation times but also to leverage the power of the pre trained transformer. For every model, we gave it 10 minutes to train (using the same machine) using a constant learning rate and the Adam optimizer. We used a train/validation/test split (70%/15%/15%), without any stratification. The results shown are from the test split. We used PyTorch as our backend.

To train our models, we decided to employ different loss functions as to optimize the modeling for different objectives.

- Our first model M_{BCE} use the BCE as the loss function, to analyze the general potential of the model.
- The second model M_{DCG} uses a SoftRank-style [19] differentiable approximation of nDCG for the loss function, to optimize for the expected nDCG. This model in theory should be better in ranking the emotions, while still giving relevance/confidence scores.
- The third model M_{MSE} will use the WMSE as our loss function, giving it the power to make more bold predictions while still penalizing aggressively wrong predictions.

In parallel, every model was trained on the 3 different emotion taxonomies previously discussed, denoted in this report as M^3, M^7, and M^{28}. We generally expect for the M^3 models to have a higher performance, due to the generalized nature of the taxonomy, followed by the M^7 models for the same reason. With that being said, we still found interesting to share the performance of the M^{28} models, as the unique taxonomy associated can share more specific emotions associated with the texts beyond sentiment and the simpler taxonomies. Additionally, as we're using annotator disagreement for the prediction directly, which is fairly uncommon (and potentially novel in emotion recognition) as discussed before, the results cannot be directly compared with ones from other articles that use this dataset, due to the big difference in the annotator aggregation. Finally, we don't expect great results in general, not only due to rig and time and constraints, but also due to the dataset quality, as discussed before.

5 Results and Discussion

After training and testing the various models using the workflow described in the previous section, we obtained the results in Table 2. From them, we can make several conclusions:

- As expected, the taxonomies with more emotions had a harder time in performing better, except when comparing the WMSE metric results; this probably means that the weight chosen wasn't penalizing lower ground truths hard enough, or this low values might be a reflection of class imbalance.
- All models are generally good at ranking the emotions, as the nDCG is high across all models. While this is not surprising in the 3 and 7 emotion taxonomies, the 28 emotion taxonomy having such high values suggest that the model is great at predicting the top emotions of the comments.
- Unexpectedly, the M_{DCG} models aren't the best ones on the nDCG metric (even though they're all similar values between them), which suggests either an issue in implementation, the unsuitability of the approximated nDCG constructed, or that the model might be overfitting if optimizing for that metric. The overall performance observed in the rest of the metrics (except in the F_1^{any}) might suggest this overfitting hypothesis.
- Surprisingly, it seems that taxonomies with lower emotions have higher MBCE and WMSE than the higher emotion

[2]We find important to note that this is preferable over a softmax activation layer, since the texts can have multiple emotions.

Table 2
The results from our models

	3 emotions			7 emotions			28 emotions		
	$M_{BCE}{}^3$	$M_{DCG}{}^3$	$M_{MSE}{}^3$	$M_{BCE}{}^7$	$M_{DCG}{}^7$	$M_{MSE}{}^7$	$M_{BCE}{}^{28}$	$M_{DCG}{}^{28}$	$M_{MSE}{}^{28}$
MBCE ↓	**0.489**	0.64	0.499	**0.286**	0.447	0.299	**0.1203**	0.509	0.124
F_1^{any} ↑	0.561	0.222	**0.688**	0.252	**0.403**	0.337	0.095	**0.346**	0.152
F_1^{conf} ↑	**0.511**	0.32	0.479	**0.279**	0.177	0.263	0.087	0.051	**0.103**
WMSE ↓	0.131	0.276	**0.118**	0.0653	0.11	**0.0649**	0.023	0.149	**0.022**
nDCG ↑	0.924	0.925	**0.929**	**0.89**	0.884	0.884	**0.801**	0.789	0.799

number taxonomy. This seems counter intuitive, but it's probably because of the big imbalance of labels per text (as most emotions should be 0). This suggests that alternative ways to model to be aware of this imbalance might be more appropriate, as well the importance of choosing a more appropriate weight function for the WMSE.

- The classification metrics pretty unstable, showing fairly different values between the different models. Regardless, from them, we can see that the models aren't as good in predicting the multiple labels directly. The M_{DCG}, weirdly enough, seems to be able to handle this better than the other emotions, when looking at the F_1^{any} metric. This might be because the model gives more confidence in the output for the emotions, as the order in the metric is more important, and there are more positive labels for the threshold we decided. This would also explain the fairly low F_1^{conf} throughout the models (including M_{DCG}), possibly suggesting that a better threshold for it should've been chosen.
- Both the M_{BCE} and M_{MSE} models perform very similarly according to the regression metrics, and both seem to be capable of predicting the emotion of the texts close to the real predictions.

Overall, the results we're in line with our expectations, and from them it seems that using and fine-tuning Distilbert is suitable for emotion recognition when taking into account annotator disagreement.

6 Conclusion

In this report, we analyzed the GoEmotions dataset with the objective of doing emotion recognition on different emotion taxonomies, these comprised of 3, 6 and 27 emotions, with the latter ones including a label for unclear. While researching and analyzing, we found some issues of the dataset, including annotator disagreement, data quality issues, and label imbalance. After aggregating the annotators based on confidence per label, we finetuned DistilBERT using different loss functions on those taxonomies. We could conclude that for most loss functions, the models mostly performed fairly similar between loss functions, but the different taxonomies had a huge impact on model performance. Ultimately, we were able create models that somewhat classify and predict the emotions in reddit comments accurately, even with heavy annotator disagreement.

In future work, there are several key directions we should explore to improve both the accuracy and reliability of emotion recognition in our models. First, we should address the label noise and cultural bias in the GoEmotions dataset. As noted during our analysis, many labels appear to be inconsistent due to annotator misunderstanding of cultural references or sarcasm. A valuable improvement would be to refine or relabel parts of the dataset using more culturally diverse annotators or by adding contextual information (e.g., surrounding conversation or media references) to help raters make more accurate judgments.

Second, although we used DistilBERT for its efficiency, experimenting with larger or more specialized language models like RoBERTa, DeBERTa, or emotion-specific transformers could help us capture more nuanced emotional signals, especially for complex cases like sarcasm or mixed emotions. We could also explore prompt-tuned or instruction-following models, which might generalize better with less fine-tuning. Additionally, data augmentation teqniques has shown to improve model performance and lift the burdern somewhat from label imbalance [5, 11, 20].

Third, to improve the model's robustness and calibration, we should consider applying calibration techniques such as temperature scaling or isotonic regression before training, especially since our evaluation showed that while confidence scores were often reasonable, some configurations still showed signs of over- or under-confidence.

Finally, we should perform a more thorough error analysis. For instance, looking into which specific emotions are most often misclassified, and under what linguistic conditions, would help us fine-tune the model and dataset further. Another option is to use model explainability tecniques like the use of shapley values [3] or LIME [17]. Future work could also expand on the hard emotion distinctions possibly incorporating label correction [12], or to incorporate fuzzy probability theory into the mix.

References

[1] 30% of Google's Emotions Dataset is Mislabeled: 2025. *https://www.surgehq.ai/blog/30-percent-of-googles-reddit-emotions-dataset-is-mislabeled*. Accessed: 2025-06-01.

[2] Ameer, I. et al. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*. 213, (2023), 118534. DOI:https://doi.org/https://doi.org/10.1016/j.eswa.2022.118534.

[3] An introduction to explainable AI with Shapley values: 2018. *https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html*. Accessed: 2025-06-01.

[4] Capstick, A. et al. 2025. Training Neural Networks on Data Sources with Unknown Reliability. (2025).

[5] Data Augmentation for Emotion Detection in Small Imbalanced Text Data: 2023. *https://arxiv.org/abs/2310.17015*.

[6] Davani, A.M. et al. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*. 10, (2022), 92–110. DOI:https://doi.org/10.1162/tacl_a_00449.

[7] Demszky, D. et al. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. (2020).

[8] Ekman, P. 1992. Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science*. 3, 1 (1992), 34–38. DOI:https://doi.org/10.1111/j.1467-9280.1992.tb00253.x.

[9] Fornaciari, T. et al. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online, Jun. 2021), 2591–2597.

[10] Galke, L. et al. 2025. Are We Really Making Much Progress in Text Classification? A Comparative Review. (2025).

[11] Imran, M.M. et al. 2023. Data Augmentation for Improving Emotion Recognition in Software Engineering Communication. *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering* (Rochester, MI, USA, 2023).

[12] Jinadu, U. and Ding, Y. 2024. Noise Correction on Subjective Datasets. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Bangkok, Thailand, Aug. 2024), 5385–5395.

[13] Oddný, L. et al. 2023. Impact of Reddit Community Culture on User Attitude Expression and Social Interaction. *Journal of Linguistics and Communication Studies*. 2, 4 (Oct. 2023), 61–67.

[14] Ovesdotter Alm, C. 2011. Subjective Natural Language Problems: Motivations, Applications, Characterizations, and Implications. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, Jun. 2011), 107–112.

[15] Plank, B. 2022. The ``Problem'' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Abu Dhabi, United Arab Emirates, Dec. 2022), 10671–10682.

[16] Plaza-del-Arco, F.M. et al. 2024. Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (Torino, Italia, May 2024), 5696–5710.

[17] Ribeiro, M.T. et al. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. (2016).

[18] Sanh, V. et al. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. (2020).

[19] Taylor, M. et al. 2008. SoftRank: optimizing non-smooth rank metrics. *Proceedings of the 2008 International Conference on Web Search and Data Mining* (Palo Alto, California, USA, 2008), 77–86.
[20] Wang, K. et al. 2024. Large Language Models on Fine-grained Emotion Detection Dataset with Data Augmentation and Transfer Learning. (2024).
[21] Weerasooriya, T.C. et al. 2021. Improving Label Quality by Jointly Modeling Items and Annotators. (2021).