# Quote-MI : Quote Multilabel classification and Interpretation

**Rashmi Vagha**
rvagha@umass.edu

**Rohit Kadappa Gundakalli**
rgundakalli@umass.edu

**Vaishnavi Panchvati**
vpanchavati@umass.edu

**Varun Tarikere Shankarappa**
vtarikeresha@umass.edu

## 1 Problem statement

In our stress-filled era, whether from hectic lifestyles, career pressures, or relationship tensions, stress is unavoidable. Many turn to wisdom or motivational quotes for encouragement. The human brain processes information differently when exposed to raw data compared to stories, sayings, or quotes. Using the right quote in the appropriate context can influence, persuade, and provide perspective and humor, distilling complex ideas into concise statements. Quotes, defined as phrases or short pieces taken from longer works or spoken by others, can be used in various technological contexts, such as text analysis, enhancing chatbots, and supporting virtual learning. For these applications, understanding the appropriate category of quotes is essential.

Sentiment analysis constitutes a fundamental task within the domain of Natural Language Processing (NLP), wherein the primary objective is to extract and discern opinions or attitudes embedded within sentences and texts. Consequently, it is often referred to as opinion mining (OM) or opinion extraction(Liu, 2012). Quotes, being a common form of textual expression, encapsulate a myriad of sentiments. While certain sentiments are overt and easily discernible, others remain implicit and pose challenges even for human comprehension.

In this project, we have developed systems utilizing various types of Large Language Models to classify quotes into multiple desired labels. We have also implemented a range of prompting techniques to enhance the straightforward interpretation of these quotes, leveraging small open-source models. Through extensive experimentation and analysis of the results, we have conducted a comparative assessment of these models to evaluate their performance and efficacy in quote classification and interpretation.

## 2 What you proposed vs. what you accomplished

While the general outline of our project remained working on Classification of quotes, we incorporated the following changes to our originally proposed plan -

- ~~Collect and preprocess dataset~~
- We proposed exploring the Multi-class Classification problem on the Quotes Dataset, however we later decided to ~~work on the more challenging problem of Multi-label Classification problem quotes dataset, and examined the performance~~
- The models that we proposed to explore were BERT (our baseline), T5 and RoBERTa, but due to compute constraints ~~we shifted to working on smaller models DistilBERT (our baseline), T5, GPT2, and RoBERTa, and examined the performance~~
- Under the interpretation task, ~~we worked on Gemma 2B model and T5 3B model to perform zero-shot, one-shot and few-shot prompting~~ as opposed to performing zero-shot, few-shot and CoT GPT2 and T5 since Gemma and t5 models are comparable in size. We decided not to perform CoT due to the lack of interpretations with CoT.
- ~~Perform in-depth error analysis to figure out what kinds of examples our approach struggled with.~~

## 3 Related work

Classification is one of the most common tasks in the field of NLP. There has been substantial research conducted on different deep learning models for NLP (Minaee et al., 2021). The pervasiveness of text classification has resulted in the necessity for understanding various models (Li et al., 2022) for classifying text and making improvements for better classification. Previous research has explored fine-tuning Transformer-based (Vaswani et al., 2017) Large Language Models (LLMs), such as Bidirectional Encoder Representations from Transformers (BERT), for text classification tasks (Sun et al., 2020). Text classification tasks have evolved from coarse-grained to fine-grained classification to capture many contexts effectively (Mekala et al., 2021). This evolution has naturally led to the exploration of multi-label classification, where each text can belong to multiple categories simultaneously, addressing the need for more nuanced and comprehensive analysis in complex datasets.

Quotes are found in almost all types of linguistic data, yet research in this area remains limited. Most early Natural Language Processing research focused on quote recommendation based on queries or conversations using cosine similarity (Tan et al., 2015). Later, similar systems were experimented with using neural network models like Convolutional Neural Networks (CNNs) (Lee et al., 2016) and Long Short-Term Memory networks (LSTMs) (Tan et al., 2018). Some recent work (Qi et al., 2022; Wang et al., 2022) has even involved Transformers, many of which use encoder-only models like BERT. Text classification using BERT generally involves fine-tuning the model with the target data and applying softmax on the CLS token to obtain label probabilities (Sun et al., 2019). Research has also explored methodologies to use encoder-decoder models for multi-label classifications (Kementchedjhieva and Chalkidis, 2023), investigating how different parts of the Transformer model can be utilized for classification.

With the rapid improvement in learning capabilities of LLMs, substantial research has been done to improve the learning capabilities of various tasks. One successful way of doing this is using In-Context Learning (ICL). In-Context methods like few-shots prompting have vastly increased the performance of the LLMs by providing few examples in the context. This enhancement relies solely on prompting without fine-tuning, meaning that none of the model parameters are altered. Despite this, it yields significant performance improvements(Wei et al., 2022). However, to get the best outcome, a suitable selection strategy for few-shots examples is necessary. We plan to use these learning capabilities for quote interpretation task.

## 4 Dataset

While there are millions of quotes available on the internet on a wide range of topics, there are however very few publicly available dataset that classify quotes into multiple labels (or tags). One of the major reasons for this can be attributed to the fact that quotes sometimes convey indirect messages that are difficult to be captured just by text summarizing or paraphrasing. As a result, performing label-classification on large number of quotes is challenging. The dataset used in the this project is the Quotes-500k dataset (Goel et al., 2018). The dataset has been created by crawling multiple websites such as GoodRead, BrainyQuote, FamousQuotesAndAuthors and CuratedQuotes. The dataset consists of approximately 500k quotes along with the author and the labels (Fig. 2).

| Quote | Author | Tags |
|---|---|---|
| A friend is someone who knows all about you and still loves you. | Elbert Hubbard | friend, friendship, knowledge, love |

Figure 1: Example of a Quote in the Dataset.

### 4.1 Data Preprocessing

The quotes have been crawled from multiple sources due of which, there are some inconsistencies in the dataset. Each quote is tagged with multiple tags. There are about 232 types of labels most of which are very specific to the quote and not very genetic, such labels are removed. The label distribution of the dataset is highly unbalanced (The most frequent tag $love$ appears 42867 times, whereas the $5^{th}$ most frequent tag $humor$ appears 14211 times), thus, a stratified sampling process is employed to ensure that the distribution of labels is comparable. This is explained in the next section.

#### 4.1.1 Quote Classification

For the purpose of Quote Classification we have used the top 5 frequently appearing labels from the dataset (Love, Life, Inspirational, Philosophy, Humor). Very long length quotes are often harder to interpret even for humans let alone computers on the other hand, shorter length failed to establish a premise for the quote, thus we decided to limit the length of each quote to between 25 and 256. After this, we employed the basic preprocessing steps such as case-conversion, language based filtering. The preprocessed dataset consists of 87941 quotes with 5 labels. The final dataset is then generated by sampling approximately 30k quotes from the latter, during the process we ensure that the label distribution is not highly unbalanced. An exact split is difficult to achieve since the quotes are multi-labeled, however, a comparable split can be achieved using some enhanced stratified sampling techniques. The final label distribution is shown in Table 1. The distribution of number of labels per quote is is shown in Table 2.

| Label | Frequency |
|---|---|
| Love | 10767 |
| Life | 11980 |
| Inspirational | 11320 |
| Philosophy | 9613 |
| Humor | 6784 |

Table 1: Label-wise Distribution

| Number of labels | Frequency |
|---|---|
| 1 | 17943 |
| 2 | 7361 |
| 3 | 2640 |
| 4 | 1483 |
| 5 | 558 |

Table 2: Multi-Label Count Distribution

#### 4.1.2 Quote Interpretation

For the purpose of Quote Interpretation we have used the top 10 frequently appearing labels from the dataset (Love, Life, Inspirational, Philosophy, Humor, Wisdom, Truth, God, Hope, Happiness, humor). The dataset was divided based on the labels and for each label, 40 quotes were shortlisted from the dataset. The quotes were manually picked for each label to ensure that each quote is interpretable.

Interpretable quote in this context refers quotes that do not require any background information, and are not very lengthy. The labels were divided among individual team members and each member selected the quotes. Once this was completed, all the selected quotes were combined and shuffled among 4 sets one per member. Each set contained 100 quotes. Each member then manually interpreted each quote, to achieve this we used sources from internet including chatGPT. We manually ensured that interpretations used from web and chatGPT accurately explains the meaning of the quote. Some of the examples are presented in Table 3.

| Quote | Interpretation |
|---|---|
| God is a witness to anything done in secret. | This quote suggests that nothing is truly hidden or secret from God; God is aware of all actions and intentions, regardless of whether they are carried out openly or in secret. |
| Every barber thinks everybody needs a haircut. | This highlights the tendency for individuals to perceive situations through the lens of their own expertise or interests. |
| the less you need, the more you live. | This quote suggests that simplifying one's desires and material needs leads to a richer and more fulfilling life, emphasizing the importance of contentment and minimalism. |

Table 3: Quote Interpretation Examples

## 5 Baselines

### 5.1 Classification

For the classification task, we opted for DistilBERT(Sanh et al., 2019), a distilled version of BERT, primarily due to its lower computational demands compared to BERT while still maintaining state-of-the-art performance for classification tasks. This choice allows for a meaningful comparison against other Transformer-based models. In traditional multi-class classification, fine-tuning BERT typically involves applying softmax over the CLS token or the last hidden state to derive individual probabilities. However, for our specific multi-label classification objective, we introduce an additional linear layer $W$ with dimensions $l_h \times l_l$, where $l_h$ represents the dimension of the last hidden state and $l_l$ denotes the number of labels. Subsequently, we apply a sigmoid function over this layer to derive *weights*, selecting label weights above a predefined threshold. This approach enables effective handling of multi-label classification tasks.

### 5.2 Interpretation

We employ the zero-shot prompting technique as our baseline measure for interpretation. This technique involves providing no additional instructions in the prompt other than the task instructions, making it the most straightforward prompting approach and a good baseline.

## 6 Approach

### 6.1 Classification

We have selected three transformer-based language models for our classification task: DistilGPT2, which is a decoder-only model; T5-small, an encoder-decoder model; and RoBERTa, an efficient variant of BERT, which serves as our baseline model. These models underwent fine-tuning using various techniques to enhance their classification performance. To facilitate model execution, we utilized the Google Colab platform, allocating 250 compute units. Furthermore, to expedite both training and inference processes, we made use of CUDA-based hardware accelerators, specifically A100 and T4 GPUs. To address session timeouts and ensure stability, we stored model checkpoints and intermediate results on Google Drive, thereby enabling seamless data persistence between sessions. Table 3 summarizes the various hyper parameters chosen for each of these models.

#### 6.1.1 RoBERTa

For classification of the tags, we fine tuned a pretrained RoBERTa model. RoBERTa, which stands for Robustly Optimized BERT Pretraining Approach is built upon BERT and optimized better by training the model longer, with bigger batches, over more data, removing the next sentence prediction objective, training on longer sequences; and dynamically changing the masking pattern applied to the training data (Liu et al., 2019). In addition to this, RoBERTa also uses byte-level BPE as a tokenizer.

In this project, we have used the pretrained 'roberta-base' which has 125M parameters and built a classification model around it using Pytorch. The architecture contains a linear layer with the input size 768 and output size 768. Next, we have a dropout layer with a dropout probability of 0.1. Finally, we have another linear layer that maps this vector of size 768 to a vector of size 5 which is the total number of target labels we have.

The loss function used was BCEWithLogitsLoss from the Pytorch nn library which computes the Binary Cross Entropy Loss between the predicted and the actual values.

#### 6.1.2 DistilGPT-2

We fine tuned DistilGPT2 for multilabel classification. DistilGPT-2 follows the same architecture as the original GPT-2 model.However, it has fewer layers and parameters compared to GPT-2, making it computationally lighter and faster. For the multilabel task,we fine tuned DsitilGPT2 by adding an additional linear layer on the top of the last token with dimensions $l_h \times l_l$,where $l_h$ is the dimension of the output layer from GPT2 and $l_l$ is the number of labels .We then applied a softmax to obtain the probabilities over different classes.We accomplished a working implementation of the model using the following libraries:

- The transformers library for fine-tuning DistilGPT2.
- The torch library for adding the linear layer on top and for creating the datasets.

The model is run on three different seeds. The model achieves an accuracy of ˜63 percent on test data.There is no considerable change in the performance compared to the baseline.The model was trained with binary cross-entropy loss and Adam optimization.

#### 6.1.3 T5

T5, or Text-To-Text Transfer Transformer(Raffel et al., 2019), is a sequence-to-sequence transformer model released by Google. For our task, we used the T5-small variant from the Hugging Face Transformers library, which consists of 60.5 million parameters. To fine-tune this model for producing labels, the quote is passed to the encoder, while the labels are passed to the decoder, with binary cross-entropy used as the loss function. For inference, we determine the labels by checking if the model's output text contains the label text. After several experiments, it was found that adding a prompt as a prefix to the quote with all the labels increased the performance.

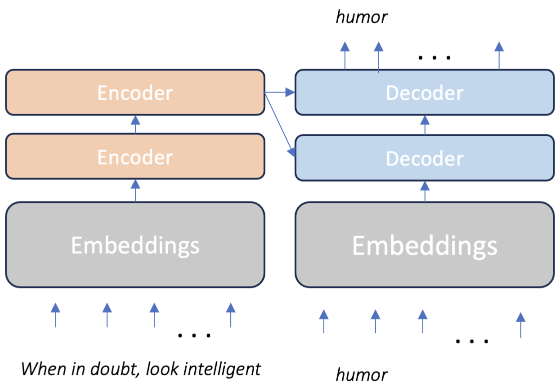


Figure 2: T5 multi-label classification

#### 6.1.4 Evaluation

The models are evaluated using accuracy.Since we have multi-label classification of the tags,we used Jaccard score to calculate the accuracy.The metric is used for comparing the similarity of predicted labels and actual labels. We also computed f1-score,precision,accuracy tags wise to compare performance between different tags.

### 6.2 Interpretation

We used T5 (3B) and Gemma (Team et al., 2024) as the models for Quote interpretation.Apart from the baseline method of zero-shot prompting,we used one shot and few shot prompting for interpreting the quotes.Figure 3 summarizes the different types of prompts with an example for each.

Gemma 2B is a decoder only transformer model which is a part of Google's latest open language models, it is trained on 2 billion parameters. Gemma advances state-of-the-art performance relative to comparable-scale (and some larger), open models (Almazrouei et al., 2023; **?**) across a wide range of domains including both automated benchmarks and human evaluation. Example domains include question answering (Clark et al., 2019), commonsense reasoning (Sakaguchi et al., 2019) and more. Thus, Gemma 2B would be a suitable model for the quote interpretation task.

#### 6.2.1 Sentence Transformer

In order to make few shot prompting more efficient, it is crucial that the quote interpretations provided as a part of the few shot prompt are related to the actual quote being interpreted. This will help the model better understand the quote by providing a context as well as confine the text generation to relevant topics. We employ the Sentence-BERT or SBERT (Reimers and Gurevych, 2019) model which is a variation of the pretrained BERT model. SBERT is specifically designed to semantically meaningful sentence embeddings which can then be compared using cosine-similarity metric. Once we have the sentence embeddings, we can find the most relevant quotes for any particular quote.

#### 6.2.2 One Shot Prompting

In one-shot prompting, we provide the model with a single example prompt along with its corresponding interpretation, and then task the model with interpreting a new quote in a similar manner.

#### 6.2.3 Few Shot Prompting

Few-shot prompting involves providing the model with three different examples of quote-interpretation pairs. The model is then asked to interpret a new quote in a similar manner, drawing from the patterns learned from the provided examples.

#### 6.2.4 Evaluation

We evaluated the interpretations generated by both the models manually using human evaluation.For each type of promting technique within a model two members of the group evaluated the interpretation and we used inter annotator agreement to reach a consensus on the final output. Interpretations were scored from 0-5 based on the following scheme:

- 0:The generated output is random and not related to the quote.
- 1: The generated output contains some key words from the quote but the meaning is not conveyed.
- 2:The generated output has an incomplete interpretation.
- 3: The generated output partially conveys the message of the quote
- 4:The generated output is very close to the actual interpretation with minor faults.
- 5: The interpretation completely conveys the essence of the quote.

| Model | Learning rate | Batch size | Epochs |
|---|---|---|---|
| DistilBERT | 2e-5 | 32 | 10 |
| T5 | 3e-4 | 64 | 10 |
| RoBERTa | 2e-5 | 32 | 10 |
| DistilGPT-2 | 1e-4 | 32 | 10 |

Table 4: Hyper parameters for classification models

| Type | Prompt |
|---|---|
| Zero Shot | Explain the meaning of Quote in 1-2 sentences.<br>Quote: happiness is part of who we are. joy is the feeling<br>**Answer:** |
| One Shot | Explain the meaning of the quote in 1-2 sentences.<br>**Quote:** happiness is holding someone in your arms and knowing<br>you hold the whole world.<br>**Translation:** This quote suggests that embracing a loved one gives a profound sense of fulfillment and joy, making the holder feel as though they possess everything they could possibly desire<br>or need in the world.<br>**Quote:** happiness is part of who we are. joy is the feeling.<br>**Translation:** |
| Few Shot | Explain the meaning of the quote in 1-2 sentences<br>**Quote:** if you work your hardest, and you do your best, then the rest is not in your control, but in god's hands.<br>**Translation:** This quote emphasizes the importance of dedication and effort in achieving goals, while acknowledging that some outcomes are beyond personal control and instead lie in the realm of fate or divine will.<br>**Quote:** do not question your ability or worthiness. god is a universe of purpose-driven balance. if you have been called to action, it is because you have it within you to rise to the challenge<br>of your calling. now, rise to it!.<br>**Translation**: This empowering message assures that if one is called to action, they inherently possess the ability to rise to the challenge, emphasizing confidence in one's capabilities and<br>worthiness to fulfill their purpose.<br>**Quote**: when my problems seem overwhelming and impossible to solve, i hold fast to these facts: god is real, he is right, and he cares.<br>**Translation** :This quote reflects a strong belief in a higher power, suggesting that during tough times when challenges appear insurmountable, one finds comfort and strength by trusting in God's existence, righteousness, and compassionate concern for individual well-being.<br>**Quote**: if everyone is against you, but god is for you. it is more than enough to keep moving.<br>**Translation**: |

Figure 3: Types of prompts

| Model | Accuracy | Size |
|---|---|---|
| DistilBERT | 61.78 | 67M |
| T5 | 60.367 | 60.5 |
| RoBERTa | 63.95 | 125M |
| DistilGPT-2 | 63 | 88.2M |

Table 5: Jaccard Similarity for each model

# 7 Evaluation

## 7.1 Classification

The metrics of the different classification models are shown in Table 4. From the table,DistilGPT-2(63%) performs marginally better than the baseline and the other models.It beats the encoder only models DistilBERT and RoBERTa by ~2 %
The comparison of different models based on training and validation accuracy is show in figure 4.The training accuracy for all the models increases to above 90 %.But the validation accuracy for the models increases slightly but becomes stagnant at about 60 %.

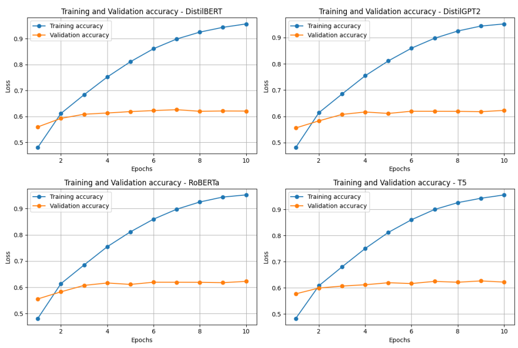


Figure 4: Training validation accuracy comparison for different models.

The loss comparison is shown in figure 5.Both T5 and DistilGPT2 have a decreasing trend in validation loss which is expected but DistilBERT and RoBERTA have an increasing trend.

We also evaluated the performance of various tags within the model. Figures 6 and 7 display the distribution of precision, recall, and F1 score for different tags for DistilBERT and DistilGPT2. The graph for DistilGPT2 indicates that the tags 'philosophy' and 'love' achieve the highest F1 scores (0.75). DistilBERT performs similar to DistilGPT2 for these two tags and also demonstrates a high F1 score(0.75) for the tag 'humor'.

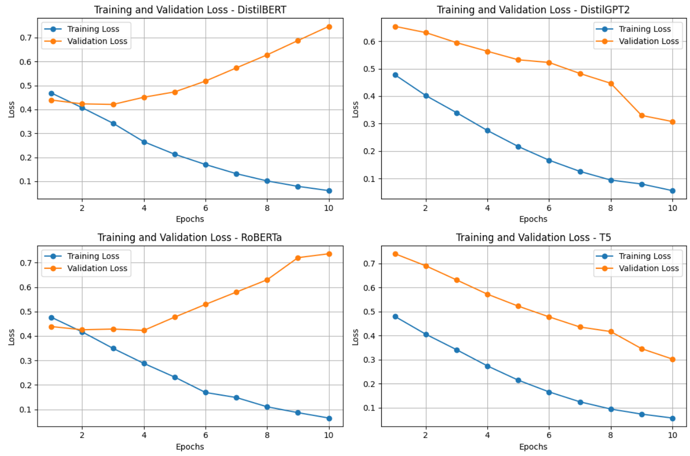


Figure 5: Training validation loss comparison for different models.

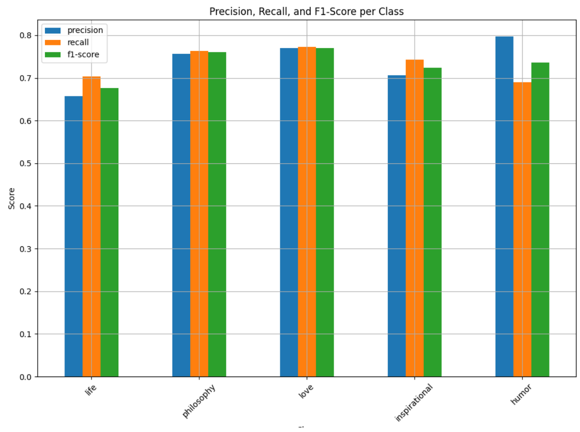


Figure 6: Comparison for DistilGPT2.

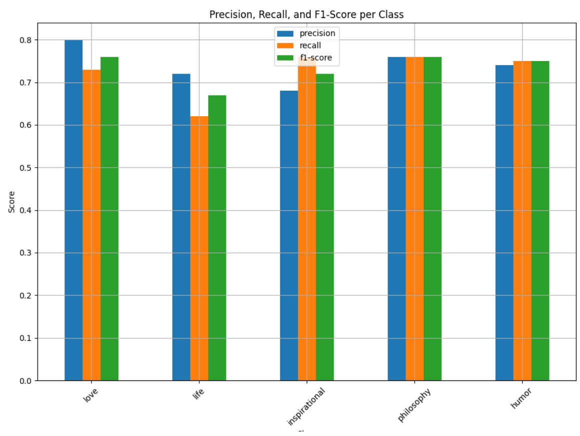


Figure 7: Comparison for DistilBERT.

### 7.2 Interpretation

We interpreted 384 quotes for each type of prompt for both the models.A combination of BLEURT (Sellam et al., 2020) and human evaluation was the original choice for the evaluation metric,but BLEURT performed poorly for scoring the interpretation.Therefore,we resorted to human based evaluation for this. The scores for the quotes are given from 0-5 for each type of the prompt.Figure 8 contains the results which is the average score of all the 384 quotes.

| Prompt | Score |
|---|---|
| Gemma Zero Shot | 4.212598 |
| Gemma One Shot | 3.76377 |
| Gemma Few Shot | 4.13910 |
| T5 Zero Shot | 0.977 |
| T5 One Shot | 1.45 |
| T5 Few Shot | 1.2461 |

Figure 8: Interpretation scores for different models

The highest score (4.2159) is achieved by the Gemma model in Zero Shot. This score indicates that the model performs well on interpretations. Surprisingly, Gemma Few Shot has a slightly lower score than Gemma Zero Shot, suggesting that increasing the number of prompts does not improve the model's interpretation. The Gemma model excels at interpreting tasks even with Zero Shot prompting. Conversely, the T5 model performs best in One Shot (1.45) but falls far short of the actual interpretation score, demonstrating poor performance in interpreting content. Changing the prompting technique to Few Shot does not significantly enhance its performance.

We also performed a tag-wise analysis for the score and averaged the score of each tag.Figure 9 shows the computed score for Gemma Zero Shot.The scores clearly show that the model performs equally well on all the tags achieving a similar score. From the scores,it is clear that amongst the tags the quotes from 'truth' tag was harder to interpret by the model. Figure 10 shows a similar analysis on T5 Few Shot. The model performs very poorly in interpreting the quotes.All the tag types have similar performance and there is no stark difference between multiple tags in performance.The tags 'life'(1.39) and 'inspiration'(1.36) get the best scores out of all the tags.

| Tag | Score |
|---|---|
| humor | 3.9756 |
| inspirational | 4.0975 |
| god | 4.375 |
| philosophy | 4.000 |
| hope | 4.2307 |
| love | 4.615 |
| wisdom | 4.684 |
| happiness | 4.305 |
| truth | 3.628 |
| life | 4.187 |

Figure 9: Scores of tags

| Tag | Score |
|---|---|
| humor | 1.215912 |
| inspirational | 1.36584 |
| god | 1.2000 |
| philosophy | 1.2000 |
| hope | 1.3333 |
| love | 1.1282 |
| wisdom | 1.2820 |
| happiness | 1.1282 |
| truth | 1.2105 |
| life | 1.394737 |

Figure 10: Scores of tags

## 8 Error analysis

We have conducted the error analysis for both the tasks i.e. classification and interpretation as follows:

### 8.1 Classification

Since we have a multi-label classification task, all the models' performances depend on predicting not just a single label but all the actual labels accurately. Few examples of where it is unable to do so are:

- "i have so much hate that it has turned into love." Here, the GPT2 model predicts 'love' whereas the actual tags are [love, humor]. The model is unable to understand the irony of too much hatred turning into love and thus it is not tagging this quote under the humor tag.
- "when god made man she was practicing."

Another challenge faced by the models is for the humor tag. It is difficult to classify this because it contains a certain amount of sarcasm. Such quotes can be complicated even for humans so it is expected that the model will not achieve a very high accuracy on this tag. The DistilBERT model is unable to classify this quote correctly.

- ”i pondered what else i should take for him. flowers seemed wrong; they’re a love token, after all. i looked in the fridge, and popped a packet of cheese slices into the bag. all men like cheese.” This is another example of the model not able to understand humor. This is for the GPT2 model.

- ”the body consists of three parts: the brainium, the borax and the abominable cavity. the brainium contains the brain. the borax contains the heart and lungs and the abominable cavity contains the bowels of which there are five: a, e, i, o, u.” We have also observed that long quotes tend to perform poorly as compared to shorter concise ones.

- ”non-duality has been the root of all religious movements on this planet. ” The issue faced by the model here is confusion between ‘philosophy’ and ‘inspirational’ tags. Both these tags are very similar in meaning and many of the quotes have an overlap between these tags. Even during manual prediction, these are the two tags which would be easily confused. There is a very fine difference and the model can only predict philosophy here and not inspirational.

### 8.2 Interpretation

For interpretation we have 3 different prompting techniques namely zero-shot, one-shot and few-shot. As expected the zero-shot prompting will not perform too well because it does not have the context or any example to generate the interpretation. In many cases, the generated text was either a repetition of the input or it was incomplete. This was especially the case with T5 zero-shot prompting. In our observation, we did not come across any grammatically incorrect texts.

- Explain the meaning of Quote in 1-2 sentences Quote: you be good. that’s your business. the rest, leave to god. that’s his business. Answer:

  you be good. that’s your business. the rest, leave to god. .... .

- Explain the meaning of Quote in 1-2 sentences

  Quote: dance. smile. giggle. marvel. trust. hope. love. wish. believe. most of all, enjoy every moment of the journey, and appreciate where you are at this moment instead of always focusing on how far you have to go.

  Answer:

  1-2 sentences dance. smile. giggle. marvel. trust. hope. love. wish. Answer:

Similar to classification issues, here the humor or puns were not easily interpreted. Certain words can have multiple meanings and for the interpretation to be accurate, it is important that the correct meaning for that word is understood.

- Explain the meaning of the quote - Quote: ’whenever he answers prayers, god usually prioritizes those by people who, instead of their mouths, have prayed with their hands and/or feet.’ Answer : This observation implies that God tends to favor those who actively work towards their prayers, embodying their faith through tangible actions rather than mere words. Quote: ’when my problems seem overwhelming and impossible to solve, i hold fast to these facts: god is real, he is right, and he cares.’ Answer : This quote reflects a strong belief in a higher power, suggesting that during tough times when challenges appear insurmountable, one finds comfort and strength by trusting in God’s existence, righteousness, and compassionate concern for individual well-being. Quote: ’the man who counts on the aid of a god deserves the help he doesn’t get.’ Answer : This quote implies that relying solely on divine intervention without taking personal responsibility may lead to disappointment or lack of assistance. Quote : is prayer your steering wheel or your spare tire? Answer : This quote suggests that prayer should not be used as a crutch or a substitute for personal effort and responsibility. Instead, it should be seen as an opportunity to connect with God and ask for guidance and support, but it is ultimately up to the individual to take action and make positive changes in their life.

Another issue while prompting these models was that they were able to generate the book name or author name of the original quote but not the meaning. This just depicts that they are pretrained on such data but not trained enough to generate the entire interpretation by themselves.

- Explain the meaning of the quote - Quote : 'what i'm looking for is not out there, it is in me.' Answer :

  The quote is from the book 'The Alchemist' by Paulo Coelho. The main character, Santiago, is on a journey to find his life's purpose. Along the way, he meets a wise old man who tells him a story about a shepherd boy who finds a treasure in the desert. Santiago is inspired by this story and decides to go in search of his own treasure, which he believes will lead him to his purpose in life.

- Explain the meaning of the quote - Quote : 'those people who think they know everything are a great annoyance to those of us who do.' Answer :

  1. This quote is from the book 'The Hitchhiker's Guide to the Galaxy' by Douglas Adams. 2. In this book, the main character, Arthur Dent, is on a journey to save the Earth from being destroyed by an alien race called the Vogons. He is accompanied by a robot named Zaphod Beeblebrox and a human named Ford Prefect. They encounter many strange and unusual creatures and situations on their journey, and at one point, they meet a man named Trillian who is an expert on the universe and its inhabitants.

## 9 Contributions of group members

- Rohit : did data processing, DistilBERT fine-tuning and Gemma-prompting, writing and annotation

- Vaishnavi : DistilGPT-2 fine-tuning and survey of evaluation metrics and compatability study, annotation and lots of writing

- Varun : assist data processing, T5-small fine-tuning and prompting, annotations and strategies for selecting demonstration (sentence model) and lot of experiments.

- Rashmi: RoBERTa fine-tuning, lot of annotation and error analysis

## 10 Conclusion

In this study, we investigated the problem of quote classification and interpretation using various transformer-based Large Language Models (LLMs). The main models explored included DistilBERT, T5, RoBERTa, and DistilGPT2 for the classification task. Our experimental results demonstrated that DistilGPT2 and RoBERTa performed similarly, with DistilGPT2 slightly outperforming the other models, achieving an accuracy of approximately $63\%$.

For the quote interpretation task, we leveraged both Gemma 2B and T5 models utilizing zero-shot, one-shot, and few-shot prompting techniques. Notably, the Gemma model exhibited superior performance in zero-shot prompting, suggesting its robust capability in interpreting quotes without the need for additional context. Conversely, the T5 model struggled significantly in generating accurate interpretations.

Our analysis also includes extensive error analysis, highlighting the models' challenges in interpreting puns, humor, and words with multiple meaning. Despite these limitations, the study underscores the potential of LLMs in automating complex NLP tasks such as multi-label classification and quote interpretation.

Some of the major challenges we faced were related to evaluating the interpretations generated by the model, which was something we anticipated. Parts of the results took us by surprise: there was no significant difference between the baseline model and the other model. This may be attributed to the fact that the sizes of the models were comparable. Furthermore, the results of few-shot prompting were also not as expected; perhaps the smaller models failed to understand the intricate details of the prompts. Given a chance to continue working on the project, exploring evaluation techniques for quote interpretation and performing classification analysis on bigger models could be good places to start.

## 11 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what

AI you used.

Yes

*If you answered yes to the above question, please complete the following as well:*

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - correct grammar and enhance coherency
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - The model gave sentences which were coherent but required further editing.

## References


Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. (2023). The falcon series of open language models.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. (2019). Boolq: Exploring the surprising difficulty of natural yes/no questions.

Goel, S., Madhok, R., and Garg, S. (2018). *Proposing Contextually Relevant Quotes for Images*, pages 591–597.

Kementchedjhieva, Y. and Chalkidis, I. (2023). An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5828–5843, Toronto, Canada. Association for Computational Linguistics.

Lee, H., Ahn, Y., Lee, H., Ha, S., and Lee, S.-g. (2016). Quote recommendation in dialogue using deep neural network. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 957–960, New York, NY, USA. Association for Computing Machinery.

Li, Z., Wang, X., Yang, W., Wu, J., Zhang, Z., Liu, Z., Sun, M., Zhang, H., and Liu, S. (2022). A unified understanding of deep nlp models for text classification.

Liu, B. (2012). *Sentiment Analysis: A Fascinating Problem*, pages 1–2. Springer International Publishing, Cham.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Mekala, D., Gangal, V., and Shang, J. (2021). Coarse2Fine: Fine-grained text classification on coarsely-grained annotated data. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 583–594, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning based text classification: A comprehensive review.

Qi, F., Yang, Y., Yi, J., Cheng, Z., Liu, Z., and Sun, M. (2022). QuoteR: A benchmark of quote recommendation for writing. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 336–348, Dublin, Ireland. Association for Computational Linguistics.

Raffel, C., Shazeer, N. M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2019). Winogrande: An adversarial winograd schema challenge at scale.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings*, page 194–206, Berlin, Heidelberg. Springer-Verlag.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2020). How to fine-tune bert for text classification?

Tan, J., Wan, X., Liu, H., and Xiao, J. (2018). Quoterec: Toward quote recommendation for writing. *ACM Trans. Inf. Syst.*, 36(3).

Tan, J., Wan, X., and Xiao, J. (2015). Learning to recommend quotes for writing. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2453–2459. AAAI Press.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. (2024). Gemma: Open models based on gemini research and technology.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Wang, L., Zeng, X., and Wong, K.-F. (2022). Learning when and what to quote: A quotation recommender system with mutual promotion of recommendation and generation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3094–3105, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.