Automatic Speech Recognition, Sentiment Analysis and Retrieval-Augmented Generation

SENTIMENT ANALYSIS & CLASSIFICATION

The first step of our work was to train a basic classification model that would be used as a baseline for further benchmarking with other models. Thus, we trained the proposed model, the Hugging Face's BERT, which is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. The configuration of this baseline model was the same as the proposed, with a maximum sequence size of 128 tokens, a learning rate of 2e-5 and a batch size of 32 samples, and a training of 3 epochs.

After obtaining the initial results, we explored the impact of different threshold values on the model's performance. The threshold determines the minimum confidence score that a predicted label must have to be considered valid. By adjusting this threshold, we aimed to find an optimal balance between precision and recall, which are critical metrics in multi-label classification. Our experimentation revealed that as the threshold increased, the precision improved while the recall decreased. This behavior is expected because a higher threshold filters out less probable labels, reducing false positives but increasing false negatives. Conversely, a lower threshold includes more labels, enhancing recall at the expense of precision.

The comparative results across various threshold values are summarized in the table below. The highest F1 score, which is the harmonic mean of precision and recall, was achieved at a threshold between 0.2 and 0.3. Specifically, the optimal F1 score of 0.49 was observed at these thresholds, indicating the best balance between precision and recall for our dataset. This threshold range ensures that the model is neither too conservative nor too liberal in label assignment, optimizing overall performance.

These findings underscore the importance of threshold tuning in multi-label classification tasks, where the trade-off between precision and recall must be carefully managed to achieve the best possible outcomes.

Comparative table of results by threshold:

Threshold	Precision	Recall	F1
0.9	0.35	0.09	0.12
0.8	0.62	0.2	0.27
0.7	0.64	0.28	0.36
0.6	0.6	0.35	0.42
0.5	0.6	0.4	0.46
0.3	0.5	0.5	0.49
0.25	0.47	0.53	0.49
0.2	0.44	0.56	0.49
0.1	0.36	0.64	0.45

In conclusion, our experimentation with different thresholds highlights the critical role of threshold tuning in optimizing model performance for multi-label classification tasks. By carefully selecting a threshold value around 0.25, we can achieve a balanced trade-off between precision and recall, leading to the highest F1 score of 0.49. This threshold allows the model to maintain a good level of precision while also capturing a significant proportion of relevant labels, demonstrating the nuanced interplay between these metrics in the context of our sentiment analysis task.

After that, we proceeded with the model tuning, testing out different configurations of the BERT model to find the one that performs better than the others. Within this process, we tried multiple values for the different hyperparameters of the model except the epochs, which we left the default value of 3, since more epochs resulted in a longer training times which conflicted with Google's Colab limitations, and fewer epochs led to under fitted models.

We tried different values for batch size, learning rate and max sequence size. About this last one, we couldn't get further than 192, since the storage space needed for the training was unaffordable by Colab's limitations. The results from the 3 different metrics (considering the macro average of each one) for the different models trained are on the table below. An important clarification is that these results are obtained using the default threshold of 0.5, further exploration of different thresholds on different models is made later.

Comparative table of results by model:

#	Max seq size	Batch size	Learning rate	Epochs	Precision	Recall	F1
1	128	32	2e-5	3	0.60	0.40	0.46
2	64	32	1e-5	3	0.59	0.35	0.41
3	192	32	2e-5	3	0.55	0.41	0.46
4	192	32	4e-5	3	0.55	0.43	0.47
5	64	64	2e-5	5	0.56	0.42	0.47

As we can see from the table, different configurations yield varying results in terms of precision, recall, and F1 score. The model with a maximum sequence size of 192, batch size of 32, learning rate of 4e-5, and three epochs achieved the highest F1 score of 0.47. This indicates a well-balanced trade-off between precision and recall. Conversely, the model with a smaller maximum sequence size of 64 and a lower learning rate of 1e-5 had a lower F1 score of 0.41, demonstrating the importance of tuning these hyperparameters to optimize model performance.

Some insightful observations are that when decreasing the maximum sequence length we did not improve the results in none of the metrics, while when we increased that hyper parameter value, we always got a better recall and a lower precision, but it led to a F1-score equal or greater than the baseline model. Increasing the learning rate was significant for the model in which we also increased the maximum sequence size, which resulted to be the best model.

Now, comparing the optimal threshold values, as well as the default value, in the different trained models, we obtain the following results:

Model #2:

Threshold	Precision	Recall	F1
0.5	0.59	0.35	0.41
0.3	0.52	0.45	0.46
0.2	0.47	0.55	0.46

Model #3:

Threshold	Precision	Recall	F1
0.5	0.55	0.41	0.46
0.3	0.52	0.48	0.49
0.2	0.47	0.57	0.49

Model #4:

Threshold	Precision	Recall	F1
0.5	0.55	0.43	0.47
0.3	0.54	0.53	0.51
0.2	0.49	0.59	0.51

Model #5:

Threshold	Precision	Recall	F1
0.5	0.56	0.42	0.47
0.3	0.52	0.46	0.48
0.2	0.45	0.57	0.48

Again, we see how the best results are obtained when dealing with a threshold of 0.2 or 0.3, with no specific preference for any of them. With our best model, we reached a f1-score of 0.51, which was a strong result, significantly stronger than the original model.

A relevant observation is that, during the threshold tuning, we also tried a threshold of 0.25, since the intuition can lead to think that it could be the optimal value since both 0.2 and 0.3 are the best performing thresholds, but it gave slightly worse results in some of the models. For example, in model 4, where we achieved the highest f1-score, with a threshold of 0.25 we got 0.5 instead of the 0.51 from the thresholds 0.2 and 0.3.

After studying the different metrics results, we finally decided to use a threshold of 0.3, since it generally lead to more stable results of both precision and accuracy. Despite 0.2 having the same f1-score, in most cases it was because the high recall compensated a low precision, while with a threshold of 0.3 it generally was due to a precision and recall around 0.53, which we believe means a more robust classification.

PROMPT ENGINEERING

Comments = <<<"It's better to say a moment like that could truly ignite her love for the game rather than putting a bit of a damper on it. I went to a destination wedding being the only single person. Promised to never put myself in that situation again. He died 4 days later of dehydration Like this just cuz of the [NAME] rhymes background raps...but dude your [NAME] is sick against [NAME] Lol dream on buddy. You’ve had enough attention today. Actually learn what your talking about helps a lot. Sorry your stuck in free roam smokin crack As an anesthesia resident this made me blow air out my nose at an accelerated rate for several seconds. Take your damn upvote you bastard. 1-2-3-4 I declare a thumb war! Dangit [NAME], you win again! Ok you get to stab me again :(">>>

Mixtral 8x7B [2-4s]

Original prompt (Summarize shortly the comments provided above. Describe general emotions.)

Answer: The comments express a range of emotions including positivity, negativity, and humor. The first comment has a positive tone, praising an experience that sparked a love for a game. The second comment has a negative tone, recounting a personal experience of feeling out of place at a wedding. The third comment is lighthearted and playful, using humor to compare names and rap music. The fourth comment is a mix of amusement and admiration, reacting to an impressive achievement. The fifth comment is sarcastic and dismissive, expressing frustration towards an argument. Overall, the comments reflect a variety of reactions and sentiments from the users.

Final Prompt

- You are a summarizer bot.
- Your task is to given a group of comments you should return a general summary of the group of comments.
- You must not summarize each comment individually, you should summarize them as a whole.

- Describe general emotions.
- Do not provide additional explanations or notes, just the summary.
- Do not use more than 60 words.
- The group of comments is inside <<<>>>.
  ###
- Here are some examples:
  ###

Answer zero-shot

The comments express a mix of emotions, including enthusiasm for a game, dislike for someone's behavior, sympathy for a loss, and lighthearted banter.

Answer one-shot

Comments share personal stories, humorous exchanges, and playful banter, with a mix of light-heartedness, frustration, and shared experiences, including wedding woes, musical appreciation, and competitive play.

Answer two-shot

Comments include personal anecdotes, humorous exchanges, and playful banter, reflecting a mix of light-heartedness, frustration, and shared experiences.

Answer three-shot

Comments include personal anecdotes, humorous exchanges, and playful banter, with a mix of light-heartedness, frustration, and shared experiences.

Answer four-shot

Comments include personal anecdotes, humorous exchanges, and playful banter, reflecting a mix of light-heartedness, frustration, and shared experiences, from a wedding experience to playful competition and casual rap mention.

Mistral 7B [1-2s]

Original Prompt (Summarize shortly the comments provided above. Describe general emotions.)

Answer: The first comment expresses disappointment that a moment didn't fuel someone's passion for a game, but instead had the opposite effect. The second comment shares a regretful experience of attending a wedding alone and promises not to do it again. The third comment contains a rap reference and a insult towards someone named [NAME]. The fourth comment is from a medical professional who finds the statement amusing. The fifth comment is a playful exchange between two people engaging in a thumb war. The overall emotions expressed in these comments range from disappointment, regret, amusement, and playfulness.

Final Prompt

- You are a summarizer bot.
- Your task is to given a group of comments you should return a general summary of the group of comments.
- You must not summarize each comment individually, you should summarize them as a whole.
- Describe general emotions.
- Do not provide additional explanations or notes, just the summary.
- Do not use more than 60 words.
- The group of comments is inside <<<>>>.
  ###
- Here are some examples:
  ###

Answer zero-shot

The comments express a mix of emotions including disappointment, regret, amusement, and friendly competition. Some comments share personal experiences, while others engage in playful banter. Overall, the tone is lighthearted and conversational.

Answer one-shot

The comments express a range of emotions from reminiscing about past experiences, frustration, humor, and friendly competition. Some comments share personal anecdotes, while others engage in playful banter and rap references. Overall, the tone is light-hearted and conversational.

Answer two-shot

The comments express a mix of personal experiences, humor, and playful banter, with emotions ranging from reminiscing about past moments, frustration, and light-hearted competition.

Answer three-shot

The comments express a mix of personal experiences, humor, and playful banter, with emotions ranging from reflection on past events, frustration, and amusement.

Answer four-shot

The comments express a mix of light-hearted banter, personal anecdotes, and humorous exchanges, with emotions ranging from amusement to frustration and surprise. Some comments share experiences of weddings, music preferences, and playful competition, while others express reactions to unexpected situations and medical experiences.

Insights

In all cases, the structured and refined final prompt overperforms significatively the original one, adding more detailed instructions, a chain of actions, as well as delimiters to follow the best prompting practices in order to achieve the desired output. The use of structured examples (from zero-shot to four-shot) refines the summarization, enhancing the models' ability to convey nuanced emotional tones and specific contexts. The consistency in core structure across different prompts highlights the models' capacity to generalize well, while the iterative refinement process underscores the critical role of precise prompt design in optimizing LLM performance for summarization tasks. However, opposite as what we initially thought, the Mixtral 8x7B with its innovative mixture of expert architecture which combines transformers with gated units, does not increase the quality of the results in anyone significatively but the original prompt, where it overperforms the Mistral 7B. The lack of specific instructions of the original prompt led the LLM to a lower performance. Since we have no established summarizing benchmark, we are basing our conclusions over the quality of the results in our own preference (human preference).