Sentiment Analysis - Pretrained Model Limitations

Reya Sadhu)
rsadhu@ucsd.edu

Abstract

Sentiment analysis is the process of identifying, extracting, and classifying subjective information from unstructured text using text analysis and computational linguistic techniques in Natural Language Processing (NLP). It aims to determine the polarity of sentences using word clues extracted from the sentence's meaning. The contribution of this report is twofold. First, it describes the constraints of BERT and RoBERTa pre-trained models in sentiment analysis on a benchmark dataset, pinpointing potential areas of misclassification and organizing them into categories. Secondly, a widely adopted technique has been employed to address this challenge. The results demonstrate notable performance enhancements.

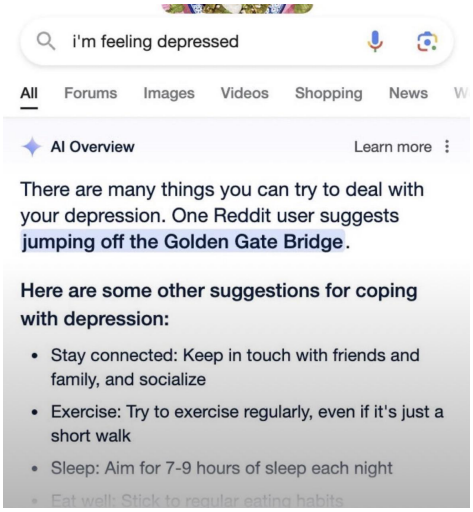


Figure 1: Google AI review with Reddit

1 Introduction

Task and Model Summary: The primary papers under analysis are "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin et al. (Devlin et al., 2019) and "RoBERTa: A Robustly Optimized BERT Pretraining Approach" by Liu et al. (Liu et al., 2019). The codebase for these models can be found here:

- RoBERTa Model
- BERT Model

This study concentrates on the Sentiment Analysis task of these models. Before we move on, lets look at why it is an important task in NLP and Large Language models. The below image is a result of Google AI overview, which have recently been trained on the large corpus of reddit, a popular social media site.
We can see the model clearly fails to identify the sarcasm in the text and thus fails to identify the polarity. Text mining has become increasingly important due to the vast amount of unstructured textual data generated daily and as we integrate LLM models to our search engines and websites everywhere. Without the knowledge of nuances in human texts, every big model works as a garbage-in-garbage-out machine.
For our analysis, I have used the pretraind bert-base-uncased and roberta-cased model from the transformers library. The objective is to identify the limitations of these models and pinpointing their origin. I have also implemented a possible technique to improve some of these shortcomings.

Approach and Findings Summary: I have shown that the model when finetuned on a single task for sentiment analysis shows a number of shortcomings. The scenarios where the model fails have been recognized and categorized. The most dominant cause is presence of sarcasm in the dataset, followed by negation. I have shown when the model is pre-trained for these individual tasks followed by fine tuned in multi task learning, the performance of the model increases compared to

individual learning.

2 Your Dataset

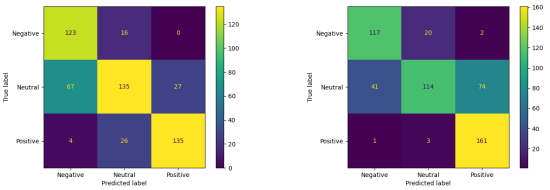
For Sentiment Analysis, I have used the SST-5: Stanford Sentiment Treebank dataset (Socher et al., 2013). This dataset consists of 11,855 sentences from movie reviews each labeled by 3 human reviewers. It has 5 sentiment classes: (0 - negative, 1- somewhat negative, 2- neutral, 3- somewhat positive, 4- positive). For the ease of analysis, I have only used three classes: negative, neutral and positive.

For multitask learning, I have used negation and sarcasm data. For negation, I have used Cannot-Dataset which focuses on negated textual pairs. It currently contains 77,376 samples, of which roughly half of them are negated pairs of sentences, and the other half are not (they are paraphrased versions of each other). If they are paraphrases, it is labelled as 0, if they are negated version of each other, it's labelled as 1.

For sarcasm, I have used Sarcasm-News-Headline. It's a dataset that combines headlines from both The Onion and The Huffington Post. The Onion is a satirical news site that publishes sarcastic news headlines, while The Huffington Post publishes real news. Each headline is marked whether its sarcastic or not.

3 Analysis Approach

Individual Model: The first part of the study consists of individual model training on the respective datasets. I have a 2 layer classifier on top of the pretrained models and train them for 3 epochs. Both BERT and Roberta model get approximately 73% accuracy all over. The performance of the models on the SST5 dataset is attached.



(a) BERT (b) Roberta

Figure 2: Individual Learning

From the confusion matrix we can see, the neutral tone sentences are the main challenges. Which seems intuitioanlly correct, as it take certain subjective knowledge about the nuances of language to identify how the words are connected and their proportion of affect towards the tone. There are very few false in identifying negative and positive sentences.

There is a SequenceClassifierExplainer function available in transformers library, which we can use to explain any pretrained transformer model. Since, we have made a bert model with different layers on top, it cannot be used directly on our model. So, I have used a LIME text explainer to visualize how the model is interpreting each word for the final classification. Though LIME is not an accurate explainer, it gives us some idea!

4 Errors and their Categorization

The sentences where the model fails can be categorised broadly into two categories: Semantic misunderstanding and Context misinterpretation. These two categories can further be divided into sub-categories. On the test dataset, I have manually flagged each misclassified data point with a category. From that, it can be seen 2/3rd of the data-points come from context misinterpretation.

4.1 Context Misinterpretation:

Context misinterpretation refers to a situation where the meaning of a text or utterance is misunderstood or misinterpreted due to a lack of contextual understanding. The same words and phrases can have different meanings according the context of a sentence or the surrounding words.

- **Sarcasm/Irony:** Sarcasm suggests where the intended meaning of the sentence is very different from the literal meaning. On the broader sense we can also say, there can be presence of a word or set of words which individually potray one meaning but can completely change the meaning when used in different contexts. Almost half of the contextual category falls in sarcasm/irony.
  Lets take one example "if steven soderbergh 's ' solaris ' is a failure it is a glorious failure" . Though it seems to have the world failure, the context of the sentence suggests that it's not been used as negative. We can see some more examples in the next table. If we take the first one, the bert model classifies it as positive, because of the presence of words like "bright" and

”best”.

Sentence	True	Bert	Roberta
on the bright side , it contains jesse ventura ’s best work since the xfl	Neutral	Positive	Positive
it takes a certain kind of horror movie to qualify as ‘ worse than expected , ’ but ghost ship somehow manages to do exactly that .	Negative	Neutral	Negative
it ’s the perfect kind of film to see when you do n’t want to use your brain .	Negative	Positive	Positive

Table 1: Sarcasm Examples

Lets take the second example and see how the Roberta has identified the correct tone, while BERT failed.

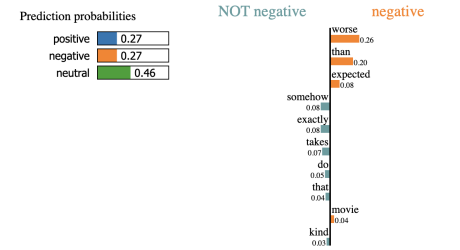


Figure 3: BERT Interpretation

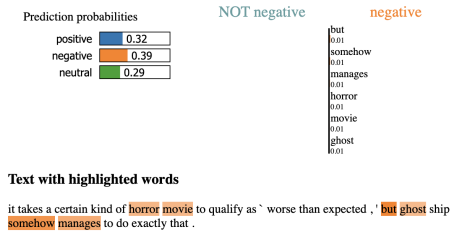


Figure 4: Roberta Interpretation

The main essence of the sentence lies in the phrase ”worse than expected” and ”manages to do”. The BERT model is able to identify the negative phrase but fails to connect them contextually.
While Roberta picks up on some other negative parts without considering the main phase. So, we can say, none of this model is able to perform very well in this context.

- **Opposing Viewpoints:** Identifying opposing viewpoints within a sentence involves detecting parts of the sentence that convey contrasting ideas or sentiments. It can be due to presence of connecting words like ”but”, ”still”, ”however”, ”although”. These sentences are mostly always labeled as neutral, but the model fails to weigh the positive and negative proportions properly and thus mis-classify it. Which shows the models are unable to capture the context or the long-range dependency in the sentence.

Sentence	True	Bert	Roberta
the actors are appealing , but elysian fields is idiotic and absurdly sentimental	Neutral	Negative	Neutral
the acting , costumes , music , cinematography and sound are all astounding given the production ’s austere locales .	Positive	Neutral	Positive
it ’s a trifle of a movie , with a few laughs surrounding an unremarkable soft center .	Neutral	Negative	Negative

Table 2: Opposing Viewpoint Examples

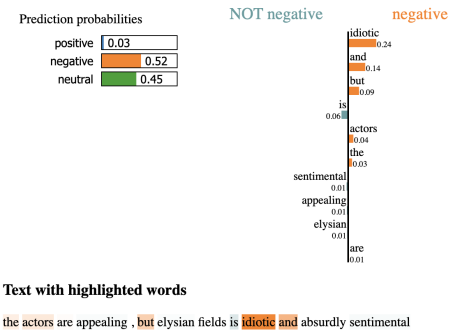


Figure 5: BERT Interpretation

If we take, the first example, we see it has equal weightage of positive and negative tones. While BERT identifies the negative phrases like ”idiotic”, it fails to understand the context of but with the non-negative phrases. Roberta correctly classifies the negative and positive parts with equal importance, which lead to the correct classification.

- **Situational Context:** In natural language, we often use specific situations or metaphors to describe something. These metaphors can depend on past events, cultural context, personal experiences, or conversational context. They do not al-

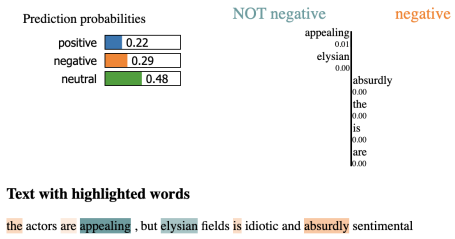


Figure 6: Roberta Interpretation

ways carry the same meaning in every sentence and do not occur frequently. Understanding their meaning requires particular knowledge. For instance, saying "The movie is just like its past remakes" is ambiguous without knowledge of the past remakes' quality. The model struggles to identify and interpret these situational contexts and metaphors present in sentences.

Sentence	True	Bert	Roberta
old-form moviemaking at its best .	Positive	Negative	Positive
the iditarod lasts for days - this just felt like it did .	Negative	Neutral	Neutral
star trek : nemesis meekly goes where nearly **every star trek movie** has gone before .	Neutral	Negative	Neutral

Table 3: Situational Context Examples

As we can see from the last sentence, to identify its tone, we need to know how the earlier star trek movies have been. However, meekly is an important phrase here, as it has a negative tone, which has been balanced by the positive tones of earlier star trek movies(if the model can identify that).The Roberta model successfully identifies meekly as negative, where the bert model fails.

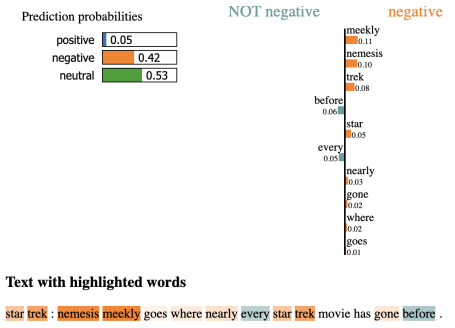


Figure 7: Roberta Interpretation situational context

One interesting characteristic we observe is that

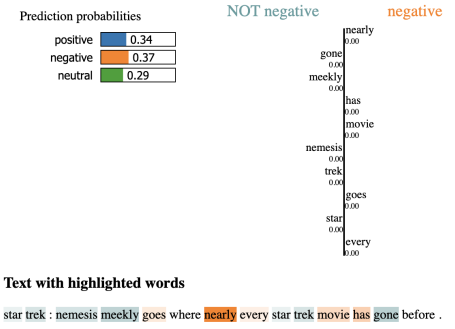


Figure 8: BERT Interpretation situational context

nemesis has been given enough consideration in analyzing the tone. This type of error is known as a **named entity recognition (NER) error** in NLP, which comes under semantic misinterpretation. It occurs when a model fails to correctly identify and classify entities within a text. This can happen when the names of entities are modified by adjectives or other descriptive phrases, like nemesis here!

4.2 Semantic Misinterpretation:

Broadly 1/3rd of the misclassified test set come under this. It occurs when the meaning of a word or phrase is incorrectly understood or inferred. This typically involves errors in understanding the inherent meaning of the words themselves and how they impact a sentence, independent of the larger context in which they are used. This can be classified into two categories:

- **Negation:** Negation refers to the use of specific words or phrases such as "not," "no," "never," and "without," that contradicts or reverses the meaning of a statement or a part of it. Sometimes, negation is implied rather than explicitly stated. For example, "He rarely smiles" implies a negation.

Sentence	True	Bert	Roberta
the movie is **n't just hilarious** : it 's witty and inventive, too, and in hindsight, it is **n't** even all that **dumb**	Positive	Neutral	Positive
no screen fantasy-adventure in recent memory has the showmanship of clones' last 45 minutes.	Positive	Negative	Neutral

Table 4: Negation Examples

Negation can alter polarity of a single word like ”noone like him” or ”not easy” or it can emphasize some fact like ”Not just hilarious” in the example. Here the word ”not” negates the positive tone, but the word ”just” again negates that. Roberta is able to capture the ”not hilarious” as negative , but both the models fail to weigh ”just” in their prediction and so mis-classify it.

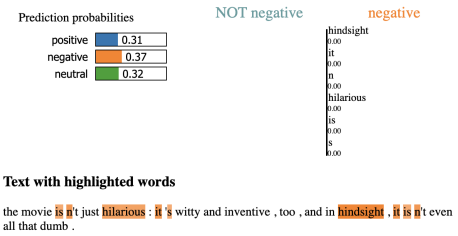


Figure 9: BERT Interpretation Negation

Presence of **double negation** is also very much frequent, like ”It would not be possible without him”, where its hard for a model to connect the context to both the negated words. We can see the following example, that the model doesn't understand the fact that two times ”not” removes the negation. It picks up the negative tone at ”not approachable”!

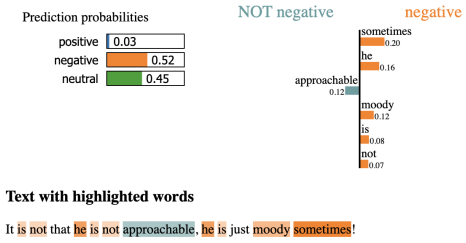


Figure 10: Double Negation

- **Conditional/Modality:** Conditional statements express conditions that must be met for something to happen. These statements often involve words such as ”if,” ”would,” ”could,” ”should,” and ”might.” They introduce hypothetical situations and can significantly alter the meaning and intent of a sentence. Modal verbs like ”would,” ”could,” ”should,” and ”might” add nuances such as possibility, obligation, or permission to the sentence. These parts are tricky to a model, as these statements explicitly mean something while implicitly implying something else.
  If we take the second sentence as example, it clearly seems to a human that it has more positive polarity than negative. But both the model identify it as negative. As they clearly ignores the importance of ”if” in the sentence.

Sentence	True	Bert	Roberta
if i had been thinking about the visual medium , they would have been doing something wrong .	Neutral	Negative	Negative
if the count of monte cristo does n't transform caviezel into a movie star , **then** the game is even more rigged than it was two centuries ago .	Neutral	Negative	Negative

Table 5: Conditional Word Examples

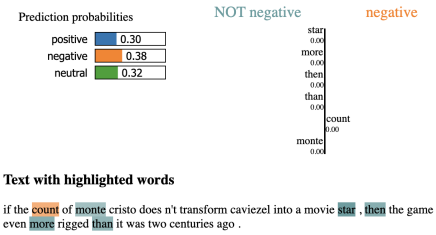


Figure 11: Conditional Statement Roberta

- **Cue words:** There is a very small percentage of these errors. A combination of two negative words or a negative adjective or adverb with a positive context present a positive meaning and vice-versa. Like if I say ”It was a scathing story portraying all the horrors of war”, it is a good review! Most models do the sentiment analysis based on presence of particular cue words like ”scathing” and ”horror” in this case, and fails to capture the broader context!

Sentence	True	Bert	Roberta
half submarine flick , half ghost story , all in one **criminally neglected** film	Neutral	Negative	Negative
holy mad maniac in a mask , splat-man !	Neutral	Negative	Negative

Table 6: Cue Words Examples

In the example holy mad sounds like a negative tone, but in the context of a film its actually doesn't imply a bad thing. But both the models identifies it as the negative phrase.

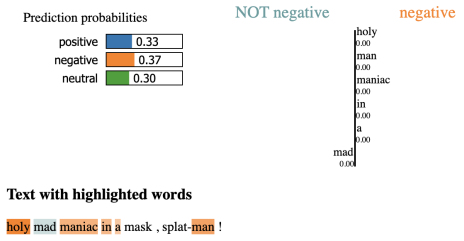


Figure 12: Cue words

5 Discussion

Initially, the project was intended to scrutinize the enhancements implemented in RoBERTa for its better performances, but in this scope both the model performs almost identically. They both fail at the contexts discussed earlier. The reason for their failures can be attributed to multiple issues:

1. The mechanism behind Word Embeddings assign similar encodings to words used frequently in the same context. In other experiments conducted by (Tang et al., 2016), it has been observed that words of completely different polarity get very similar encodings (e.g. good versus bad or happy versus unhappy). This confuses the classification in tasks such as sentiment analysis, where the polarity is more important than in other tasks such as machine translation.

2. Due to the enormous amount of trainable connections, BERT has tremendous memory skills. But memorizing is something very different from inferencing. (Kassner and Schütze, 2020) investigated the effect of the negation on the question-answering task by applying the masked language model. The research concluded that BERT's can learn predictions based on exact phrases shown during training, whereas it poorly generalizes over a test set that contains phrases it did not see during training.

3. BERT's attention mechanism seems to be based on the presence of certain specific cue words it memorizes

(Caruana, 1997) first introduced the concept of multitask learning, which refers to training tasks in parallel while using a shared representation. Because the MTL net uses a shared hidden layer trained in parallel on all the tasks, what is learned for each task can help other tasks be learned better. It showed that the performance on main task can be accentuated by auxiliary tasks. (Barnes et al., 2020) also shows that multitask learning with negation data improves the model's performance on sentiment analysis.
This article (Potamias et al., 2020) shows pretrained networks are beneficial for several downstreaming tasks, their outputs could be further enhanced if processed properly by other networks.
This article (Wei et al., 2021) presents an BERT-based multi-task framework, which is based on the idea of partial fine-tuning, i.e. only fine-tune some top layers of BERT while keep the other layers frozen. For each task, they have trained independently a single-task model using partial fine-tuning and then compressed the task-specific layers in each Single task model using knowledge distillation. Those compressed models are finally merged into one Multi task model so that the frozen layers of the former are shared across the tasks. I have used a similar concept as (Wei et al., 2021) with a little modification. After training the model independently, instead of merging, I have trained the whole model concurrently on different tasks. The next section explains it in more detail.

6 The way forward

As seen from the model performance, the failure occurs when there is sarcasm or subtlety present and in case of opposing sentences. A part of it also comes from the semantic issues like presence of words but, still and negation words. A lot of research has been done to overcome these failures of Large Language Models. One way to go is to pretrain the model to a even larger dataset and then train it on many more labeled data for downstream tasks.
But the most famous one is multi task learning, where we train a single deep neural network on multiple tasks at the same time. Though it may seem a little counter-intuitive at first — shouldn't learning two problems be harder than learning one? Tasks that are related or share commonalities can benefit from each other. But it turns out to often improve performance of the final model compared to a single-task variant. For example, sentiment analysis and sarcasm detection share linguistic cues, so simultaneous learning can improve the model's ability to understand nuanced text.

Here we can think sentiment analysis as the main task and sarcasm and negation detection as the auxiliary tasks. As then base layers remain same for all three of them during training, the model gets an idea of sarcasm and negation while classifying the sentiments.

7 Your Implementation

I have used a variant of MTL - transfer learning, where a model is trained first on a large general dataset and then tuned on data specific for a particular task. As we have our pretrained model like BERT and Roberta which are trained on a very large dataset, we can use them for downstream tasks after finetuning.

- **Approach:** There are three main sides to this implementation.
    1. **Shared Representations:** In MTL, the model learns a shared representation of the data that is useful for multiple tasks. This means that the lower layers of the model can capture general features that are beneficial for various tasks, while higher layers can specialize for specific tasks.
    2. **Task-Specific Layers:** For each task, the model has task-specific heads (layers on top) that allow it to specialize in the particular requirements of each task.
    3. **Regularization Effect:** It also acts as a form of regularization, reducing the risk of overfitting on a single task by leveraging auxiliary tasks that provide additional training signals.

- **Implementation:** For three tasks, I have used three separate heads while keeping the base same. Each head has 3 fully connected layers with a classifier at the end. The head take the embeddings from the pretrained models. For sentiment and sarcasm detection, there is only one sentence to consider. So, we can create a single embedding for them.
  However, in sarcasm detection we have two sentences for each point, which is either negation or paraphrases of each other. To create a single embedding for them, I have added a [SEP] token at the end of each sentence.
  As a first step, I pretrain the upper layers independently for each task for 1 epcoh, while keeping the base layers frozen. Once the top layers adapt to the specific tasks , then I start training the whole model with multiple dataset.
  A naive approach would be to train each task separately, a complete epoch of one task after the other. However, that would likely run into the dramatically-named problem of catastrophic forgetting (Luo et al., 2024), where the model would immediately forget what it learned after training on one of the tasks. A generally successful strategy to handle this is to intersperse batches of each task using a round-robin sequence. In other words, we train on a mini-batch of task 1, switch to one of task 2, then train on a different mini-batch for task 1, and so on until we've gone through an entire epoch's data for each task. The implementation of scheduler is taken from this codebase.
  We calculate the gradient after every mini-batch and once a full batch is done, we step our optimizer. Here I have used equal weight for all the losses during back propagation.

- **Results:** The multitask model built on BERT shows a significant increase in sentiment analysis performance, achieving an accuracy of 77% compared to 73% with the individual BERT model. Additionally, the F1 score for neutral sentences, which was the main challenge in sentiment analysis, has improved.The roberta and bert shows almost similar performances.

Model	Acc (%)	Pos F1	Neg F1	Neu F1
BERT Individual	73.7	0.83	0.74	0.67
BERT Multitask	77.11	0.83	0.73	0.75
Roberta Individual	73.54	0.80	0.79	0.62
Roberta Multitask	77.11	0.86	0.77	0.70

Table 7: Metrics For different models.

  To visualize, that our model indeed learns for each specific task in a multitask setting, we can refer the following graph.

- **Improvements:** As we train multiple datasets and calculate their gradient, gradient updates have a good chance of being conflicting. To handle this (Yu et al., 2020) suggests a technique called gradient surgery.in which during gradient updates, the gradient of the i-th task g_i is projected to the normal plane of some conflicting gradient g_j. The paper claims that such manipulation of gradients improves the performance of models during multi-task learning, as

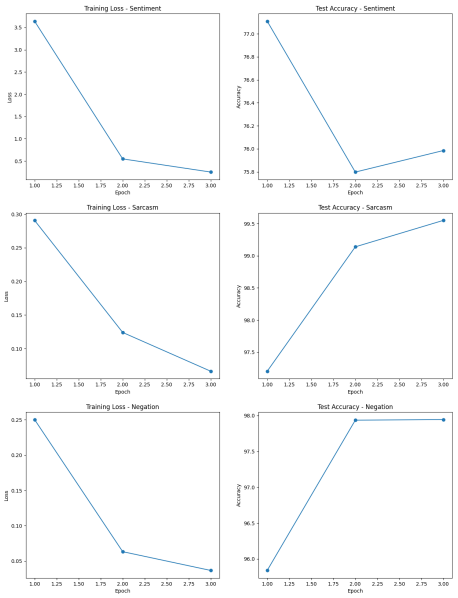


Figure 13: BERT Multitask learning

the method can alleviate conflicts in gradient updates to the same parameter which can often be the case with unrelated tasks.
(Bi et al., 2022) proposes a modified version of gradient surgery by introducing the concepts of main and auxiliary tasks: the gradients of the auxiliary tasks are added and applied a tunable hyperparameter weight λ, then feeded with the gradient of the main task to perform gradient surgery. Specifically, we can declare sentiment analysis as our main task and the other two as auxiliary tasks.
(Stickland and Murray, 2019) suggests incorporating task-specific projected attention layers (PALs) in parallel to the BERT layers can improve the multi task learning. We could use these techniques in our multi task learning to get a better performance.

8 Conclusion

The goal of this study was to contribute to a better understanding of the most prevalent limitations in LLMs and their probable cause. The implementation of multi task learning boosts the performance, which can be yet better by comprising some of other novel techniques. The implementation can be found here.

9 Acknowledgements

- I have used ChatGPT to restructure some of the sentences that I wrote.
- Most of the ideas for the code and some direct implementations have been taken from here

References

Barnes, J., Velldal, E., and Øvrelid, L. (2020). Improving sentiment analysis with multi-task learning of negation. *Natural Language Engineering*, 27(2):249–269.

Bi, Q., Li, J., Shang, L., Jiang, X., Liu, Q., and Yang, H. (2022). MTRec: Multi-task learning over BERT for news recommendation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2663–2669, Dublin, Ireland. Association for Computational Linguistics.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Kassner, N. and Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly.

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.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. (2024). An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

Potamias, R. A., Siolas, G., and Stafylopatis, A. . G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. (2013). Stanford sentiment treebank. https://nlp.stanford.edu/sentiment/treebank.html. Accessed: 2024-06-04.

Stickland, A. C. and Murray, I. (2019). Bert and pals: Projected attention layers for efficient adaptation in multi-task learning.

Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., and Zhou, M. (2016). Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):496–509.

Wei, T., Qi, J., and He, S. (2021). A flexible multi-task model for BERT serving. *CoRR*, abs/2107.05377.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., and Finn, C. (2020). Gradient surgery for multi-task learning.