GPT and BERT for Sentiment Analysis: a review

Abstract

This paper examines the performance and accuracy of two NLP/Machine learning models, BERT and GPT-3.5, when inferring the sentiment behind short texts. Both models are extremely large and powerful models that have been trained in very large datasets. Based on the results we obtained using a synthetic dataset, this study demonstrates that GPT has a higher accuracy than BERT when it comes to sentiment analysis.

Introduction

Sentiment analysis is a technique in Artificial Intelligence (AI) that combines Machine Learning (ML) and Natural Language Processing (NLP) in order to extract the sentiment behind a text. It involves evaluating the sentiment communicated in a message and determining whether it is positive or negative. It is a powerful tool that can be used for many things, such as analyzing customer feedback. This is also called opinion mining.

Related work

Task and dataset

In this paper, we will work through different text examples to classify them using NLP models to determine whether they are positive or negative. This is a type of classification (binary classification since we have two classes: positive and negative).

Binary classification is the process of sorting elements into two different categories called classes.

Here are some examples from a textual dataset for sentiment analysis, with their associated (true or groundtruth) labels:

- “Wow, this book was the worst read ever! Do not recommend it!” → Label: Negative
- “This movie was the BEST I have seen in my 70 years of life! What a masterpiece!” → Label: Positive

For this task, we have generated a synthetic dataset using GPT-3.5 to try with our models. We have 110 sentences in the dataset. We have used this prompt to generate our dataset:

> *Produce 110 examples for sentiment analysis.*
> *Examples are categorized as either positive or negative. Produce 55 negative examples and 55 positive examples.*
>
> *Use this format for the examples:*
> *Text: <sentence> Label: <sentiment>*

Here are 10 sentences from our generated dataset:

> Text: "I just finished reading an amazing book that left me feeling inspired and uplifted." Label: Positive
>
> Text: "Spending time with my loved ones always fills my heart with joy and gratitude." Label: Positive
>
> Text: "The sunset today was absolutely breathtaking, reminding me of life's beauty." Label: Positive
>
> Text: "I received a promotion at work today, and I couldn't be happier!" Label: Positive
>
> Text: "Attending the concert last night was an incredible experience; the music was mesmerizing." Label: Positive
>
> Text: "I got stuck in traffic for hours today, making me late for an important meeting." Label: Negative
>
> Text: "The food at the restaurant was terrible; it was cold and tasteless." Label: Negative
>
> Text: "My laptop crashed, and I lost all my unsaved work; I'm beyond frustrated." Label: Negative
>
> Text: "Dealing with customer service was a nightmare; they were unhelpful and rude." Label: Negative
>
> Text: "I received a rejection letter from my dream university, and I'm devastated." Label: Negative

These are the sentences that are going to be used to test our models. You can find the complete dataset in the following spreadsheet: sentiment analysis - DATASET .
Positive labels are represented by a 1 and negative labels are represented by a 0.

Methods

In order to complete this task, we are going to be using two NLP/ML models: BERT and GPT-3.5.

Short for Bidirectional Encoder Representations from Transformers, BERT is a model developed in 2018 by Google that can be used for sentiment analysis, text generation, summarization, question answering, and translation. BERT is an NLP model based on Transformer architecture (especially the encoder part). A Transformer is an artificial intelligence (especially NLP) model based on neural networks.
In order to test the BERT model, we used the pipeline module from HuggingFace, which is one of the most efficient ways to use pretrained models for inference tasks such as sentiment analysis and text and image classification.

GPT-3.5, short for Generative Pre-Trained Transformer, is a general-purpose language prediction model capable of using information to generate content. We tested the GPT-3.5 model through the Chat-GPT web interface which currently uses this model.

In the next section, we will examine and discuss the results of our experiments with these two models.

Results

Here, we show then discuss the results that were obtained using these models.

Models	**Well Classified Sentences**	**Accuracy of the model**
BERT	105	95.45%
GPT	110	100%

We can see that GPT-3.5 has obtained 100% accuracy. This is most likely because the sentences were generated by Chat-GPT, meaning that it has already seen these sentences and therefore Chat-GPT knows the sentiment of these sentences.

Here are the sentences that weren't well classified by BERT:

I found a $100 bill on the street. (Positive, but classified as Negative)
I received a handwritten letter from a long-lost friend. (Positive, but classified as Negative)
I mastered a difficult skill after months of practice. (Positive, but classified as Negative)
I passed my driving test on the first attempt. (Positive, but classified as Negative)
I was stood up on a date. (Negative, but classified as Positive)

BERT struggles with some positive sentences. For example, in the sentence “I mastered a difficult skill after months of practice”, we can assume that BERT sees the “difficult skill” part and assumes that it has a negative connotation. We can say the same thing for “long-lost”. For the sentence with “driving test”, maybe it has seen in the past example where “driving test” was associated with negative things.

Conclusion and Future Work

In this experiment, we compared the performance of GPT-3.5 and BERT for sentiment analysis. In order to do so we created a synthetic dataset using GPT-3.5 of 110 sentences with 55 positive sentences and 55 negative sentences. Based on the results we obtained, we can conclude that GPT is more accurate than BERT, having correctly identified the sentiment of all 110 sentences. BERT on the other hand identified 105 out of the 110 sentiments correctly, making the accuracy 95.45%. 4 out of the 5 sentences where it identified the sentiment incorrectly were positive, while only 1 was negative. One possible explanation for this was that some words used in the sentences that were classified incorrectly had a negative connotation despite the overall sentiment being positive. GPT-3.5’s accuracy is most likely due to two main reasons. The first one is that Chat GPT-3.5 is more recent than BERT and was trained with a larger dataset, making it more powerful. Furthermore, the sentences were all generated by Chat GPT-3.5, meaning that it was already familiar with these sentences.
In the future, we would like to test both models using a dataset unknown to both models in order to be able to make a better comparison between the two different models.

References

Shivanandhan, Manish. “What Is Sentiment Analysis? A Complete Guide for Beginners.” *freeCodeCamp.Org*, freeCodeCamp.org, 30 Sept. 2020, www.freecodecamp.org/news/what-is-sentiment-analysis-a-complete-guide-to-for-beginners/

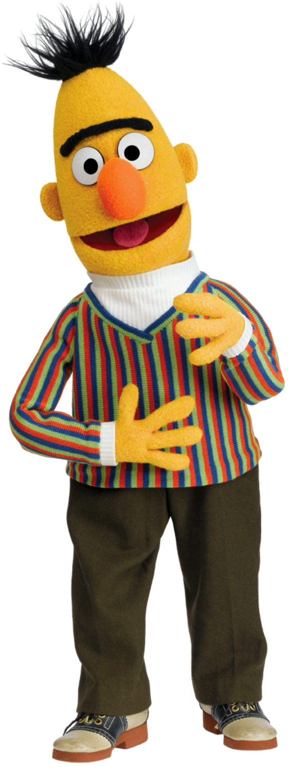
Bert 101 - state of the art NLP model explained. BERT 101 - State Of The Art NLP Model Explained. (n.d.). https://huggingface.co/blog/bert-101

What is GPT? GPT-3, GPT-4, and more explained. Coursera. (n.d.). https://www.coursera.org/articles/what-is-gpt

“🤗 Transformers.” 🤗, huggingface.co/docs/transformers/index. Accessed 9 Feb. 2024.

Quick tour. (n.d.). Huggingface.co. Retrieved February 9, 2024, from https://huggingface.co/docs/transformers/quicktour

“What Is Natural Language Processing?” *IBM*, www.ibm.com/topics/natural-language-processing. Accessed 9 Feb. 2024.



BERT