Model Performance Report: Evaluating Models for Semantic Search

Mishary Alotaibi

May 2023

1 Overview

This report aims to assess the performance of three selected models: all-mpnet-base-v2 and all-MiniLM-L6-v2, in the context of the semantic search for Twitter Bios. The process of evaluating each model was by sampling 100 Twitter bios from 20000 Twitter bios, and then we embedded each bio using the given model. Then, we used Pinecone to get the top-k similar bios. Lastly, we used a simple metric to evaluate each model. The highest-scored model was all-MiniLM-L6-v2.

2 Mythology

As mentioned in the overview, we have sampled five times, and each sample had 100 Twitter bios, all randomly sampled from 20000 Twitter bios to avoid bias while evaluating the models.

After we sampled, we preprocessed each Twitter bio in order to have condensed information without unnecessary words, such as stop words. Then, we embedded each bio using the given model and then inserted them into Pinecone in order to do cosine-similarity.

For our metric, we displayed the top 5 similar Twitter bios to the first Twitter bio in the given sample. Then, within the top 5, we would see the top Twitter bio. If the top Twitter bio is similar to the input Twitter bio, based on our perspective, then we would give the model +1. If not, but at least one of the Twitter bios within the top 5 is close, then we would give the model +0.5. If neither, then the model will receive +0. After we calculated the samples' scores, we would calculate the overall score. The overall score can be calculated by: $\frac{A}{Total}$

A: the cumulative score is based on how close the model is based on our perspective.

Total: how many samples that we used, which is, in our case, 5.

3 Models

3.1 all-mpnet-base-v2

According to the Hugging Face website, the model was developed as part of the project: "Train the Best Sentence Embedding Model Ever with 1B Training Pairs." Hence, the model was trained on 1 Billion training tuples, making it one of the largest models for semantic search. The embedding dimensions for this model are 768, which is the typical size of a sentence transformer.

When we used the model on the five different samples, the model had an overall score of 0.5. And here is the breakdown:

First dataset = +1

Second dataset = +0.5

Third dataset = +0

Fourth dataset = +0.5

Fifth dataset = +0.5

Overall score = $\frac{(1+0.5+0+0.5+0.5)}{5} = 0.5$

3.2 all-MiniLM-L6-v2

According to the Hugging Face website, the model was developed using a self-supervised contrastive learning objective and fine-tuned on a 1 billion sentence pairs dataset using the pre-trained nreimers/MiniLM-L6-H384-uncased model. The embedding dimensions for this model are 384.

When we used the model on the five different samples, the model had an overall score of 0.7. And here is the breakdown:

First dataset = +1

Second dataset = +0

Third dataset = +0.5

Fourth dataset = +1

Fifth dataset = +1

Overall score = $\frac{(1+0+0.5+1+1)}{5} = 0.7$

4 Discussion

In this study, we evaluated the performance of two models, all-mpnet-base-v2 and all-MiniLM-L6-v2, using Twitter bios as the dataset. However, while the findings provide valuable insights, several aspects warrant discussion.

Firstly, the subjective nature of the problem introduces a certain degree of bias into the evaluation process. The metric employed to evaluate the models, which depends on our perspective of similarity between Twitter bios, may not entirely reflect the models' capability. This is inherently due to the variability in human judgment - what seems similar to one observer might not be seen as similar by another. Thus, the score assigned to each model is somewhat subjective and could differ with a change in the evaluator.

Secondly, we exclusively used Twitter bios as our dataset. While Twitter bios offer a wealth of information, they have a unique structure and language style that may not fully represent other text forms. Thus, the model's performance on this particular type of data might not directly translate to other data types. It would be beneficial to incorporate different types of datasets in future evaluations to evaluate the models in a more comprehensive and generalized manner.

In conclusion, while our current methodology and findings provide meaningful initial insights, it is crucial to account for these factors to conduct a more thorough, objective, and generalized evaluation of the models. For more insights, please visit the third referenced website, which has diverse datasets and metrics.

5 References

1. all-mpnet-base-v2

2. all-MiniLM-L6-v2

3. https://www.sbert.net/docs/pretrained$_{m}odels.html$