Semantic Relations Analysis using BERT-Extracted Features

Amin Rezaei

Computer Science BSc - Amirkabir University of Technology

November 27, 2021

Abstract

Deep Language Models in particular Transformers have demonstrated impressive capabilities in recent years, and it appears that the models have reached a satisfactory level of comprehension. This implies that the features extracted from them are most likely rich. To investigate the transformers, we plan to analyze semantic relationships and classify texts using BERT-extracted features.

1 Language Modelling

1.1 Language Models

Language modeling is concerned with the representation and distribution of words capable of recognizing words and their relationships to one another, and generally proceeds with the approach of guessing the next words in a sentence or text or guessing empty words. The main goal of this model is to learn how to properly represent words, sentences, and language structure.

RNNs and, in particular, LSTMs were used to model language prior to the appearance of Transformers and Self-Attention; after that, Transformers appeared, which had a large number of parameters compared to the predecessor models. BERT, whose Large variant had 340M parameters, is an example of a successful transformer [2].
These models are typically trained on large corpora crawled from the internet, with a portion of sentences masked and the model expected to correctly predict the masked parts. Gradient descent is typically used for training, and after convergence, training is terminated. Training these models is typically expensive and time-consuming due to the large number of parameters and corpus sizes.
These models are not just for completing masked sentences; it has been demonstrated that the models achieve good language understanding and can be used as a Language Model; thus, the extracted representations can be used for many tasks, directly or with fine-tuning on special data and architectures [3]. Many other larger models, such as OpenAI's GPT-2 with 5.1B and GPT-3 with 175B parameters, have been presented [4].

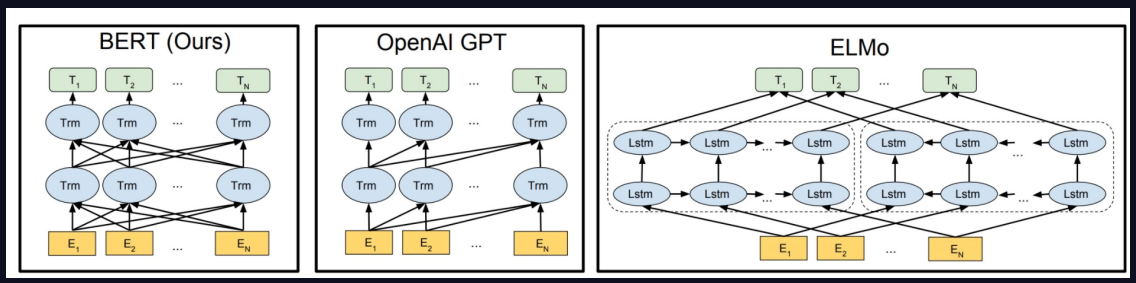

Figure 1: Architectures of Some Transformers as Language Models [5]

1.2 BERT Architecture

As previously stated, this model is trained on large corpora with the goal of predicting the masked portions. BERT is made up of multiple encoder layers stacked on top of each other, with a task-specific head on top. Each variant has its own embedding sizes and encoder count. Figure 2 depicts the overall architecture.

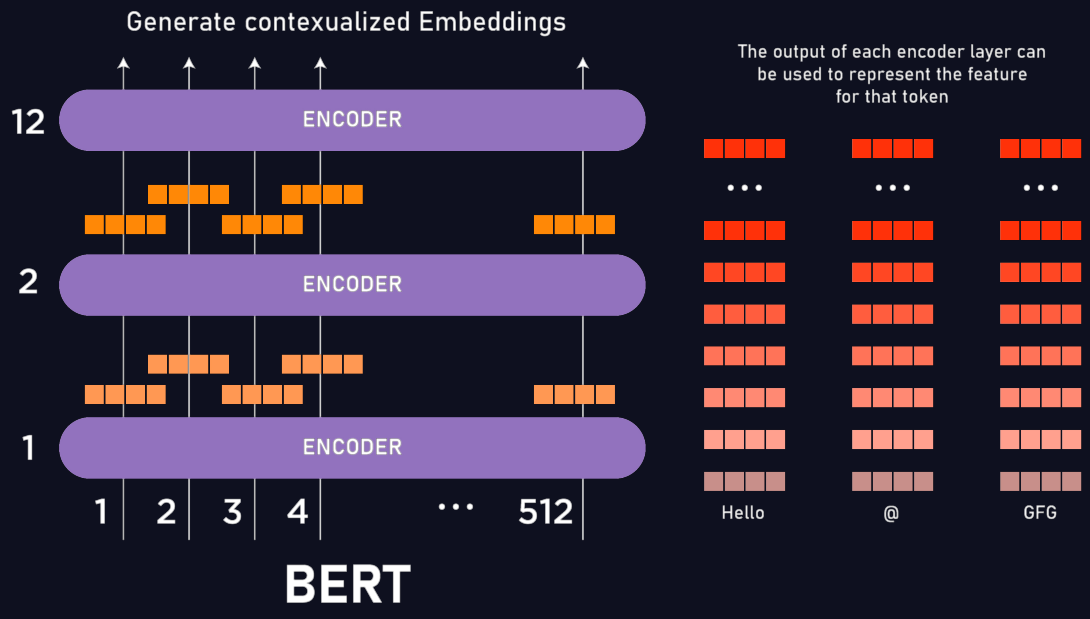

Figure 2: BERT Architecture [6]

1.3 ALBERT

As the name "A Lite BERT" implies, this model is a lite version of the BERT Architecture that is identical to BERT in terms of input/output but includes optimizations that reduce the model's size by 30% and improve performance in many tasks. The following optimizations have been applied: [7, 8]

- Decomposition of Embedding Matrix to smaller Matrices
- Cross-Layer parameter sharing and smaller Encoder segments.
- Use "Sentence Order Prediction" instead of "Next Sentence Prediction"

2 The goal and implementation

Dataset We will use the BBC News dataset, which contains 2225 news posts from five categories from the years 2004 and 2005 [9].

Goals We intend to implement following tasks:

- Extract Feature Embeddings from ALBERT
- Train a Classifier Network on extracted features
- Perform Semantic Search on Posts

Approach Our approach is to use the ALBERT features directly, with no Fine-Tuning on ALBERT itself.

2.1 Extract Feature Embeddings from ALBERT

We'll use the pre-trained ALBERT from huggingface [10]. We will use the nltk library [11] to tokenize sentences and words. For each input token, the model generates a vector of 768 items, which the pooler aggregates to one vector of the same size for each sentence. The output of the model's pooler will be used as the feature representation. There will be two sets built:

- Sentence Features Set

  This set tokenizes each post's sentences and uses the post's category as the sentence class, giving us "features $\rightarrow$ category" for each sentence.

- Post Features Set

  This set is created by mean-pooling each post's sentences and representing them as a single vector of 768 items.

2.2 Train a Classifier Network on extracted features

We create a Neural Network called CatNet that is trained on the Sentence Features Set and attempts to learn how to categorize sentences.

Figure 3: CatNet Architecture

Executor, GlobalManager, ExecutionEvents, PreprocessingDataLoader classes are helper classes designed to ease training and validating the models. For displaying progress, tqdm library is used.

For training, we use Categorical Cross Entropy as the loss function and Adam as the optimizer with a learning rate of 0.001.

To prevent overfitting, the Early Stopping technique is used, which stops and selects the best model if there is no progress after 12 consecutive epochs of training. The model is trained for 150 epochs using 80% of the data for training and 20% for testing. Figure 4 depicts the training procedure. As shown, training is terminated after 60 epochs using Early Stopping to avoid overfitting with use of validation data.

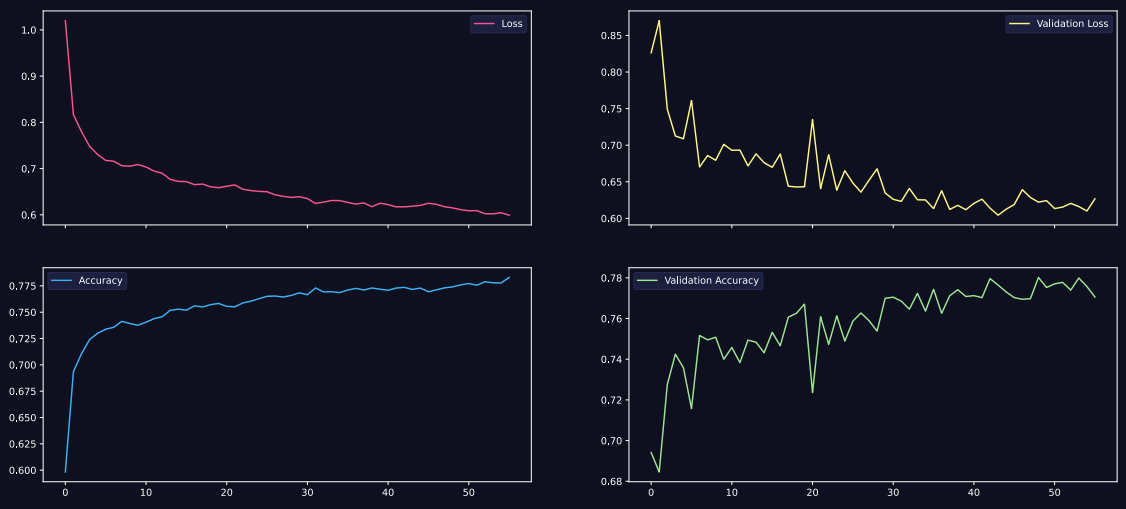

Figure 4: CatNet's training procedure

Model reaches 78.3% accuracy on training set and 78.4% on the test set which is acceptable and good performance.

Baseline We use a decision tree classifier as Baseline model. Sklearn's implementation is used [12]. Trained decision tree reaches 99% accuracy on training set and 61.45% on the test set which is far less than CatNet's performance.

We categorize some out of the dataset posts with the models to observe the performance (sources: CNBC, UsaToday) :

What McDonald's minimum wage raise says about fast-food franchise future

McDonald's is among fast-food franchises to raise wages in a tight labor market and plans to reach an average of $15 an hour by 2024 at all company-owned restaurants. Competition for workers is intense and food franchises like McDonald's and Chipotle are competing with retailers like Amazon, Walmart and Target, McDonald's CEO Chris Kempczinski noted in a recent CNBC Evolve interview.

CatNet: Business

DecisionTree: Business

Biden presses Putin to disrupt cybercriminals in Russia as U.S. grapples with latest ransomware attacks

President Joe Biden pressed Russian President Vladimir Putin in a phone call Friday to take action to stem recent ransomware attacks from Russia-based groups.
The call came just days after a massive new cyberattack by the group REvil, believed to be based in Russia.

CatNet: Politics

DecisionTree: Business

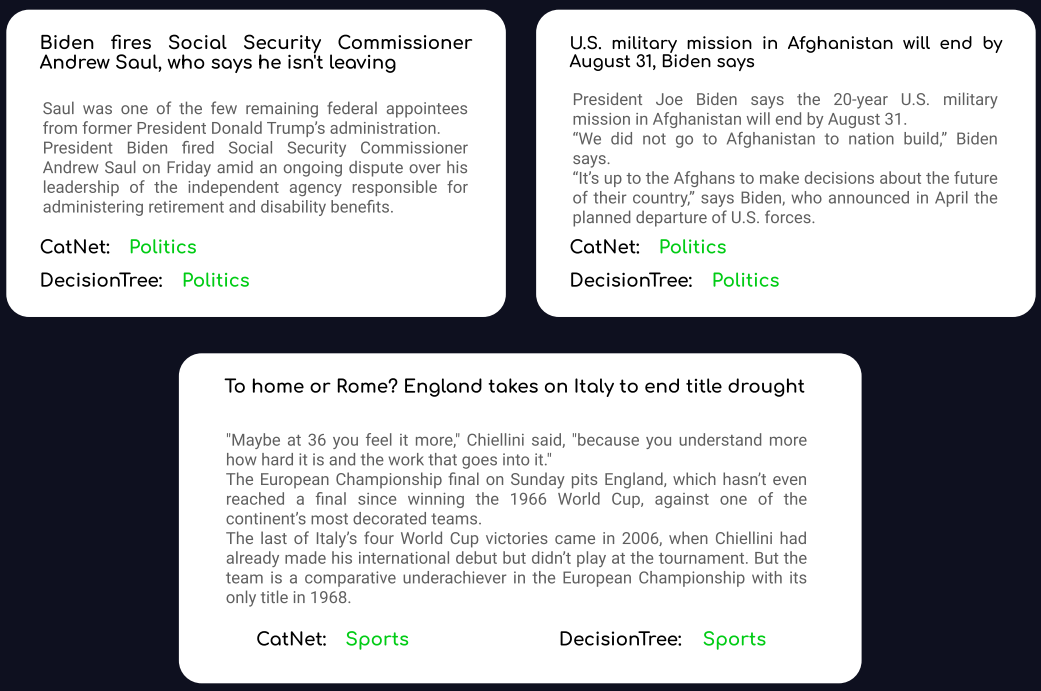

Figure 5: Results on out of the dataset posts

CatNet correctly categorizes all samples, whereas Decision Tree makes one error. Both models performed satisfactorily, but the performance difference would be more noticeable in a wider test.

2.3 Semantic Search on Posts

We look for similar posts by locating the K-Nearest Neighborhoods in the Post Features Set to see if semantically similar concepts have a lower distance in the learned space. Sklearn's implementation is used [13].

To find similar posts, we use one post's features in the dataset and two other sentence features that are not in the dataset (as if they were user queries) as inputs.

The following is the outcome of the search:

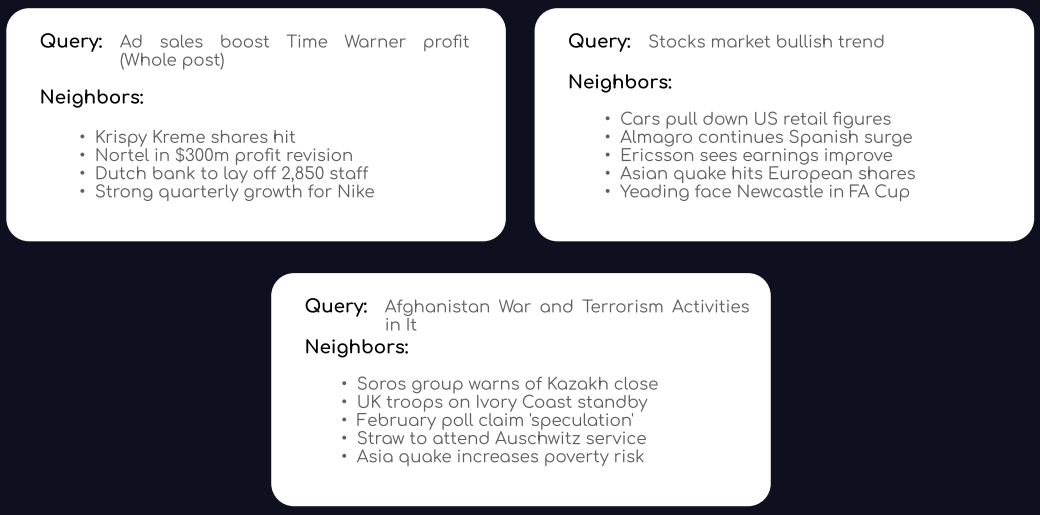

Figure 6: Nearest Neighbor Search Result

We can see that this method works extremely well in the first case for locating similar posts. For the latter two, the found posts are acceptable given that there isn't much information in the query and the extracted features aren't as rich as post features. The results are semantically similar, and the approach's overall performance is acceptable. This technique can be combined with fuzzy search methods to improve performance.

3 Conclusion

We discovered that ALBERT's extracted features were rich and powerful. We achieved acceptable results without any model fine-tuning. For improved results, stronger models, better feature selection, and even fine-tuning ALBERT can be considered. Also, as seen in the semantic search section, the features capture semantic relations well. A small but powerful search engine can be created by combining semantic search capabilities with keyword and fuzzy search techniques, eliminating the need for users to enter specific keywords. Strong knowledge bases or graphs can also be built from features that can be used in a variety of tasks.

Complete implementation is provided in the following repository: https://github.com/AminRezaei0x443/Semantic-BERT

References

[1] A. Mohammadpour. Data mining course lecture and slides.

[2] Yashu Seth. BERT Explained. https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[4] NVIDIA. OpenAI Presents GPT-3, a 175 Billion Parameters Language Model. https://news.developer.nvidia.com/openai-presents-gpt-3-a-175-billion-parameters-language-model.

[5] Analytics Vidhya. Demystifying BERT: A Comprehensive Guide to the Groundbreaking NLP Framework. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/.

[6] BERT Struct Picture. https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/.

[7] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[8] ALBERT Explained. https://www.machinecurve.com/index.php/2021/01/06/albert-explained-a-lite-bert/.

[9] BBC News Dataset. https://www.kaggle.com/pariza/bbc-news-summary.

[10] Pretrained ALBERT Model. https://huggingface.co/albert-base-v2.

[11] NLTK Library. https://www.nltk.org/.

[12] Sklearn DecisionTreeClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

[13] Sklearn NearestNeighbors. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html#sklearn.neighbors.NearestNeighbors.