

Research Paper Recommendation System using Transformer Model

Snehal S. Nayse¹ and Mrs. Pratiksha R. Deshmukh²

^{1,2}College of Engineering, Pune (COEP)/ Department of Computer Engineering and IT, Pune, India
Email: snehalsn20.comp@coep.ac.in, dpr.comp@coep.ac.in

Abstract—Searching for relevant material on the internet has become a big difficulty with the growing number of internet users and the large volume of data available on the internet. Traditional search engines return a lot of irrelevant results in response to a user's search query, wasting a lot of time and effort because the context and semantics of the user's request are not examined. This emphasizes the requirement for an embedded approach in the search engine, allowing semantic web search. Keyword based search engines fetch online pages by comparing the tokens in the user query to the tokens in the web documents. This strategy has several flaws. To overcome these obstacles, the suggested system employs semantic search using a Sentence transformer model. The system that has been proposed can speed up data retrieval and improve data searching to provide relevant outcomes to the user. The proposed model uses the arXiv dataset which contains an archive of research papers and using BERT and FAISS, it gives relevant research paper titles as the output to the query. For semantic retrieval of data, the model behaves as a vector-based search engine. This vector-based model will save time and give efficient results to the user's query.

Index Terms—semantic search, Bidirectional Encoder Representations from Transformer (BERT), data retrieval, Facebook Artificial Intelligence Similarity Search (FAISS), vector-based model, natural language processing.

I. INTRODUCTION

Traditional keyword-based search engines do not provide accurate results to the user query. Keyword search engines frequently have difficulty with: long queries, such as a paper abstract or a paragraph from a blog, complex user queries, or terms with many meanings. A semantic search is a search approach that improves precision by recognizing the cause for the search as well as the contextual magnitude of phrases as they appear. Two principles govern semantic search: search intent and semantic meaning. Search engines must analyze content based on both of these aspects in order to read natural language more properly or contextually. Vector-based (also known as semantic) search engines avoid these issues by applying state-of-the-art language models to identify a numerical representation of text queries, indexing them in a high-dimensional vector space, and calculating how similar a query vector is to the indexed pages. Users will have an easier time finding answers due to Semantic Search. The proposed algorithm for the search engine can improve the quality of results by using the indexing process, vectorizing, and ranking methods. The user query results would be more accurate than previous models. The model will predict the most relevant research papers according to the user's query from a large set of research papers available.

II. PRELIMINARIES

A. Natural Language Processing

NLP will play a key role in bridging the gap between human interaction and digital information. When technologies like deep learning and machine learning are combined, computers can analyze human language as speech data or text and 'understand' its precise meanings, such as the speaker's or writer's intention and sentiment. We need this type of technology in semantic searching. While humans are great at using conversational context and previous knowledge to deal with an inherent ambiguity in words, this may not be the case with search engines, especially when it comes to searches that aren't in the dictionary. It is possible to train a Language Model on a huge corpus of documents using the most recent findings from NLP research. Following that, the model can represent documents using their "semantic" content. This includes, for example, the ability to look for papers with semantically comparable content.

B. BERT

Researchers at Google Research presented Bidirectional Encoder Representations from Transformers as a Natural Language Processing Model in 2018. It has achieved state-of-the-art accuracy on several different NLP and NLU tasks when it was proposed. BERT builds on the typical Transformer design of multi-head attention between all term pairs in a text sequence, but it goes far more profound. BERT-Large is the most advanced version, with 24 Transformer layers, each with 1024 hidden dimensions and 16 attention heads. BERT undergoes pre-training deep-bidirectional representations by conditioning all layers on both right and left contexts. Pre-training architecture and fine-tuning architecture are shown in Figure 1 and Figure 2 below. This is done by two unsupervised prediction techniques as given below:

Masked Language Modeling: This refers to the prediction of the word by hiding the original words intentionally. By masking some fraction of the input tokens randomly and then predicting only tokens that are masked, the Masked Language Model allows us to train the bidirectional model.

Next Sentence Prediction: Many essential downstream types of tasks, such as Natural Language Inference and Question Answering, rely on an understanding of the relationship between two text sentences, which language modeling does not directly represent. Two sentences are given as input here. The authors of the research pre-train a binarized next sentence prediction task that can be easily built from any monolingual corpus in order to train a model that understands sentence relationships.

The benefit of BERT is that it may be fine-tuned for downstream tasks like Phrase Pair Classification, Question Answering, Sentence Tagging, and so on, to offer embedding or encoding that retains the characteristics of the message sentence.

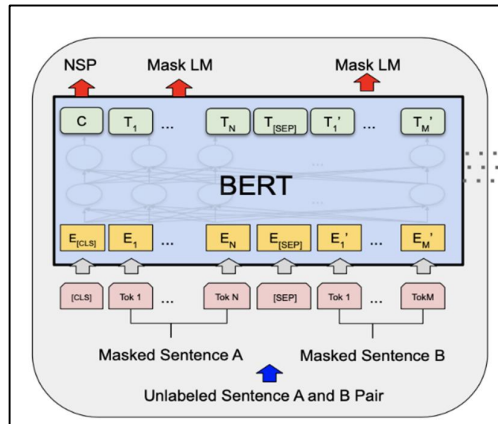


Figure 1. BERT Architecture (Pre-training)

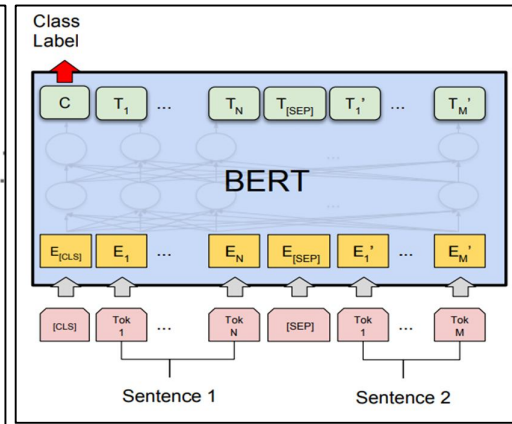


Figure 2. BERT Architecture (Fine-tuning)

C. FAISS

FAISS (Facebook AI Similarity Search) is a library that lets developers search for similar embeddings of multimedia documents rapidly. It overcomes the drawbacks of typical query search engines, which are specialized for hash-based searches, by providing more scalable similarity search features. Developers can use FAISS to search multimedia materials in ways that are inefficient or impossible to achieve with traditional database engines (SQL). FAISS strives to provide cutting-edge performance across the board.

FAISS includes algorithms for searching in sets of vectors of arbitrary size, as well as evaluation and parameter adjustment routines. On the GPU, some of the most valuable algorithms are implemented. FAISS is written in C++ and has a Python interface as well as CUDA GPU capabilities. This model needs efficient similarity search and FAISS provides us the same.

III. LITERATURE REVIEW

Ref. [1] the author, Subham Roy examines Semantic Web, Semantic Search, and different types of Semantic search engines, as well as the differences between keyword base search and Semantic Search, and also the benefits of Semantic Search. The paper also provides a brief overview of the history of semantic search and its potential in the future. [2] In this Paper, on keyword-based web techniques and Semantic techniques for retrieval of information, the author, Vajenti Mala has used six web search engines, among which three are keyword-based (Google, Wikipedia, and Yahoo), while the others are semantic web-based (Bing, DuckDuckGo and Hakia) for relevant information using the precision ratio formula, where 3 is the most important for Effective measurement. A comparison of keywords and semantic search engines' performance is shown here. It implies that semantic searching is essential for good information retrieval. Ref. [3] Bzar Khidir Hussan has written a paper that provides an overview of the differences between different Semantic-Search Engines so that anyone interested can conduct further research. The details are provided that are very useful. Ref. [4] Mr. Koroteev has given us a detailed paper about the model's mechanism of operation, the key areas of applicability to text analytics tasks, comparisons with similar models in each task, and a description of some proprietary models. The paper provides us with a detailed study of BERT and its architecture. It also shows the process of construction of text models that are domain-specific. It provides data from several dozen original scientific studies published in recent years that drew the most attention in the scientific community and were organized by the author. Ref. [5] Jacob Devlin has introduced us to BERT, a new language representation model that is mentioned in this paper, which is a short form for Bidirectional Encoder Representations from Transformers. The detailed architecture of the BERT model is given in the paper. Pre-training and fine-tuning of BERT are explained here. Demonstration of the importance of bidirectional pre-training for language representations is shown. BERT makes use of masked language models to provide deep bidirectional representations that have been pre-trained. Ref. [6] Manish Patel has created a semantics search engine that uses neural networks and BERT-embeddings to find query and answer results. The outcome shows the arrangement of documents in order of importance from most important to least important. The project uses a Deep neural network and BERT technique to obtain accurate results. Ref. [7] Anup Anand Deshmukh, has used concepts like NLP and BERT to carry out this project. This paper outlines two approaches to the TREC 2020 News Track background linking challenge. The main target of this project is to offer a list of relevant articles to the reader in order to comprehend the query article's context and obtain background information. The author's first method focuses on constructing an appropriate search query by combining weighted keywords derived from the query document and extracting them using BM25. The second method makes use of SBERT's ability to learn semantic search over the corpus. Ref. [8] describes an experiment to compare the performance of three semantic vector types - ELMo, random indexing, and GloVe as well as two different classification architectures—linear regression and multi-layer perceptrons for the task of identifying the authors with eating disorders from the writings they post on a discussion platform. The classifier must analyze texts produced by writers in the order in which they were published and identify authors who are at risk of developing eating disorders as soon as possible. Contrary to the expectations, the author could not find a clear advantage in using the recently popular contextualized ELMo vectors over the commonly used vectors. Ref. [9] Singh, A.K.; Shashi offers a vectorization strategy based on word targets to detect unifiable news items. They establish a structure for locating news on current subjects and hashtags. Then, depending on the trending subjects, they undergo a multi-document summarization of news. Previously, they used text clustering to break down the news corpus relevant to each trending topic into smaller, unifiable groups. They have used Document Understanding Conferences 2004 benchmark dataset, k-means algorithm, and purity measure to evaluate the performance of several text vectorization techniques, including bag of word representations with scores of tf-idf, document embedding, and word embeddings. Ref. [10] by combining LDA with Topic2Vec, the author created document-topic vector representations, and the document representations were performed that are based on the topic vectors and document vectors obtained by Doc2Vec that is trained. They apply their method to document classification.

IV. PROPOSED METHODOLOGY

We are randomly looking through billions of web pages in order to extract and exchange data from multiple sources. While people are quite good at using conversational context and previous knowledge to deal with an inherent ambiguity in words, this may not be the case with search engines, especially when it comes to searches that are out of the vocabulary searches. The answer to this problem is semantic search. Using the latest insights from NLP research, it is possible to train a Language Model on a large corpus of documents. We are going to build a vector-based search system. Vector-based (also called semantic) search systems avoid issues like complex queries, words with a dual meaning, or queries that are long by applying language models to identify a numerical representation of text queries, indexing them in a high-dimensional vector space, and calculating how similar a query vector is to the indexed pages. The proposed model is shown in Fig. 4. There are various tasks that need to be carried out for efficient results of the searched query. Firstly, an arXiv dataset is selected to work upon. After selecting the dataset, it needs to be loaded into a data frame. Then, quickly peek into the columns and data provided. The next step is to perform data cleaning and pre-processing steps. Data pre-processing which is one of the most significant steps in text analytics has been performed on the dataset. The purpose is to remove any unwanted words or characters or columns or even sentences that are unnecessary. The dataset chosen for this project contains abstracts that do not need too much data cleaning. After this, vectorization of documents is performed using the BERT model. Sentence Transformers is a framework that makes dense vector representations of words, paragraphs, and images simple to compute. The models use transformer networks such as BERT, RoBERTa, etc to overcome the lack of performance in a variety of tasks. Text is embedded in vector space in such a way that comparable texts are close together and may be detected quickly using cosine similarity. After vectorization, the indexing of documents is performed using the FAISS library. FAISS has algorithms that search through collections of vectors of any size, including those that are too large to fit in RAM. Lastly, the results are checked whether the proposed model is providing us with the relevant results. The evaluation metric we are using in this model is the nDCG@5 score and Precision. These algorithms and libraries make the model different from the existing ones. The model gives research paper titles related to the user's query. A simple flow of the proposed model is presented in Fig. 3. below.

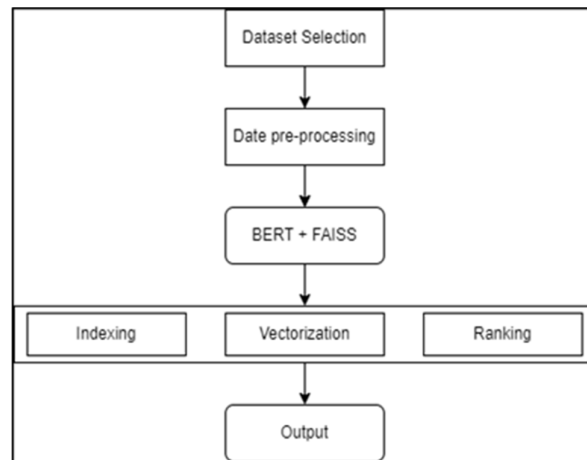


Figure 3. Flow Diagram

The dataset that we have chosen for the proposed model is the arXiv research papers dataset. This dataset is a large repository of 1.7 million articles with various relevant features like id, authors, title, versions, license, abstract, categories, etc. This dataset is a replica of Arxiv data and the original dataset is of 1.1TB and it is still growing. There are ten attributes in the dataset. One of the data cleaning steps performed is the removal of some of the columns that were unnecessary.

These are some of the features present in the dataset:

id: ArXiv ID (to access the research paper), submitter: who submitted the research paper, authors: Authors of the research paper, title: title of the research paper, categories: Categories or tags, versions: A version history.

The proposed model uses libraries like pandas, pickle and torch. Cuda is installed in the system as the model needs GPU. When vectorizing documents using Transformers, it is advised that we use a GPU.

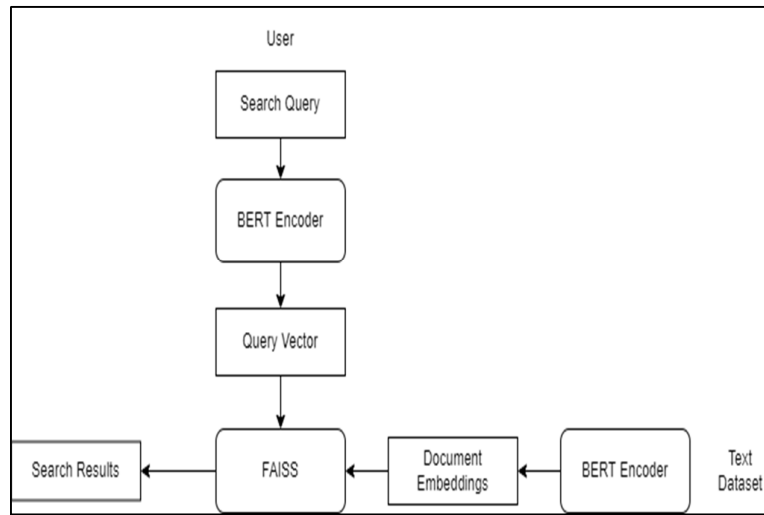


Figure 4. Proposed model

V. CONCLUSION AND FUTURE WORK

Finding semantic similarity between two sentences has always been a difficult task in natural language processing. For its ability to do many NLP tasks better than prior techniques, BERT, announced by Google in 2018, signals the start of a new age in NLP. Research papers are very important for the study of different topics in today's world. This proposed model will be able to give all the relevant research papers according to the user's given query. As we are using BERT in this model, the system will be able to achieve promising results. The vector-based model helps in solving difficulties by applying state-of-the-art language models to create a numerical representation of text queries, indexing them in a top vector space, and calculating how similar a search vector is to the indexed documents. The FAISS that is going to be used in the model makes it very fast than other similarity search systems. The proposed work will be useful to the academic area. This model will turn out to be very helpful to all the researchers and students who are in search of various research papers for their study purpose.

REFERENCES

- [1] Debabrata Barik Surajit Goon Subham Roy, Akshay Modak. An Overview of Semantic Search Engines. International Journal of Research and Review, October 2019
- [2] K.Lobiya Vajenti Mala. Semantic and Keyword Based Web Techniques in Information Retrieval. International Conference on Computing, Communication and Automation (ICCCA2016), IEEE 2016, April 2016.
- [3] Bzar Khidir Hussan. Comparative Study of Semantic and Keyword Based Search Engines. Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 1, 106-111 (2020), March 2020.
- [4] Koroteev M.V. BERT: A Review of Applications in Natural Language Processing and Understanding. arXiv preprint arXiv:2103.11943 (2021), March 2021.
- [5] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, May 2019.
- [6] Manish Patel. TinySearch- Semantics based Search Engine using Bert Embeddings. arXiv preprint arXiv:1908.02451 (2019), August 2019
- [7] Udhav Sethi Anup Anand Deshmukh. IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles. arXiv preprint arXiv:2007.12603 (2020), July 2020.
- [8] Fano, E.; Karlgren, J.; Nivre, J. Uppsala University and Gavagai at CLEF Erisk: Comparing word embedding models. In Proceedings of the Working Notes of CLEF 2019 Conference and Labs of the Evaluation Forum (CLEF 2019), Lugano, Switzerland, 9–12 September 2019; Volume 2380
- [9] Singh, A.K.; Shashi, M. Vectorization of Text Documents for Identifying Unifiable News Articles. Int. J. Adv. Comput. Sci. Appl. 2019, 10.
- [10] Peng, H.; Wang, J.; Shen, Q. Improving Text Models with Latent Feature Vector Representations. In Proceedings of the 2019 IEEE 13th International Conference on Semantic Computing (ICSC), Newport Beach, CA, USA, 30 January–1 February 2019; pp. 154–157.