International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 6 June 2023, pp: 618-622 www.ijaem.net ISSN: 2395-5252

Semantic Search Semantic search – Search based on the context and retrieve documents based on meaning and not on the keywords.

Shreyas Bhosle , Vaishnavi Kanhegaonkar , Shailesh Galande
Computer Department, Pimpri Chinchwad College of EngineeringPune,India

---
Date of Submission: 14-06-2023 Date of Acceptance: 24-06-2023

ABSTRACT: A cutting-edge method for information retrieval, semantic search employing BERT's algorithm makes use of deep neural networks to comprehend the meaning of text. A pre-trained language model called BERT, or Bidirectional Encoder Representations from Transformers, has the ability to interpret text in a manner that is comparable to how people comprehend language. By considering the context and meaning of words and sentences rather than merely matching keywords, this algorithm produces more accurate search results. Search engines may provide consumers more relevant results by using BERT's algorithm for semantic search, which enhances the search process overall.
Keywords:-semantic search, BERTalgorithm, sentence embeddings, similarity, search engines.

I. OVERVIEW

A cutting-edge method for information retrieval based on the usage of deep neural networks to comprehend the meaning of text is semantic search employing BERT's algorithm. Semantic search goes beyond traditional keyword-based search engines in that it considers the context, purpose, and meaning behind the words used in a query. Conventional keyword-based search engines focus on precise matches of words and phrases to provide search results.

A pre-trained language model called BERT, which stands for Bidirectional Encoder Representations from Transformers, is able to interpret text in a manner that is comparable to how people comprehend language. With its transformer design, the deep learning system known as BERT excels at tasks requiring language comprehension.

II. OVERVIEW OF MODEL USED

- **Berts Algorithm**

In 2018, Google researchers developed BERT, also known as Bidirectional Encoder Representations from Transformers, a powerful pre-trained natural language processing (NLP) model. It is based on a neural network design known as the Transformer architecture, which was released in 2017 and aims to improve the performance of sequence-to-sequence applications like machine translation.

As a pre-trained model, BERT was given a tremendous amount of text data to use in teaching it general language patterns and correlations. Because to this pre-training, BERT may be used to many different NLP tasks, negating the need to start from scratch with fresh models. One of the key innovations of BERT is bidirectional training, which allows the model to make predictions while taking into account the context of a word or phrase. The capacity to consider the context of a word or phrase's location in the text beyond its immediate surrounds was limited by earlier models that only used unidirectional training.

- **Sbert Algorithm**

Sentence-specific A variation of the BERT (Bidirectional Encoder Representations from Transformers) approach is BERT, often referred to as SBERT or Sentence-BERT. BERT was first created for language modelling and classification applications, but SBERT is mainly focused with developing high-quality sentence embeddings that properly capture the semantic meaning of the text. To do this, SBERT use a siamese network design, which examines word pairings and establishes their degree of similarity..

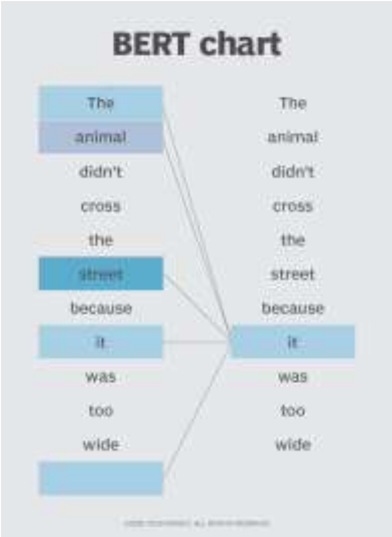

Fig 1.1 Berts Algorithm

III. LITERATURE SURVEY

Ref no	Author/Year	Research Paper Name	Publisher Name	Proposed Solution
[1]	Nils Reimers and Iryna Gurevych/ 2019	Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks	IEEE	The research suggests a techniquefor creating sentence embeddings using a Siamese BERT architectural variation, which entails creating sentence embeddings that capture semantic similarity using two identical BERT models .
[2]	SUNILKUMAR P and ATHIRA P SHAJI	A Survey on Semantic Similarity	IEEE	In this paper, the semantic similarity of text texts is surveyed. An essential job in natural language processing is semantic similarity (NLP). Information retrieval, text categorization, question answering, and plagiarism detection all make extensive use of it.
[3]	R.GUNA and Eric Miller	Semantic search	IEEE	In this work, we offer an application named "Semantic Search," which is based on these underlying

				technologies and intended to enhance conventional online searching.
[4]	Jacob Devlin,Ming-Wei Chang, Kenton Lee,Kristina Toutanova/2019	BERT: Pre-training ofDeep BidirectionalTransformers forLanguage Understanding	IEEE	Using two unsupervised learning tasks—masked language modelling and next sentence prediction—this paper's suggested method entails pre-training a deep bidirectional transformer model on a large corpus of text data.

IV. IMPLEMENTATION

As a result of searching for meaning as well as keywords in text, semantic search increases the likelihood that a user will discover the information they're looking for. Semantic search has significant implications; for instance, it would enable developers to search for code in repositories despite their lack of programming knowledge or ability to foresee the appropriate terms. Also, you may apply this strategy broadly to a variety of items, such as audio and video material as well as things we haven't yet considered. As seen here, the key concept is to merge text and the object we're looking for (code) in the same vector space.

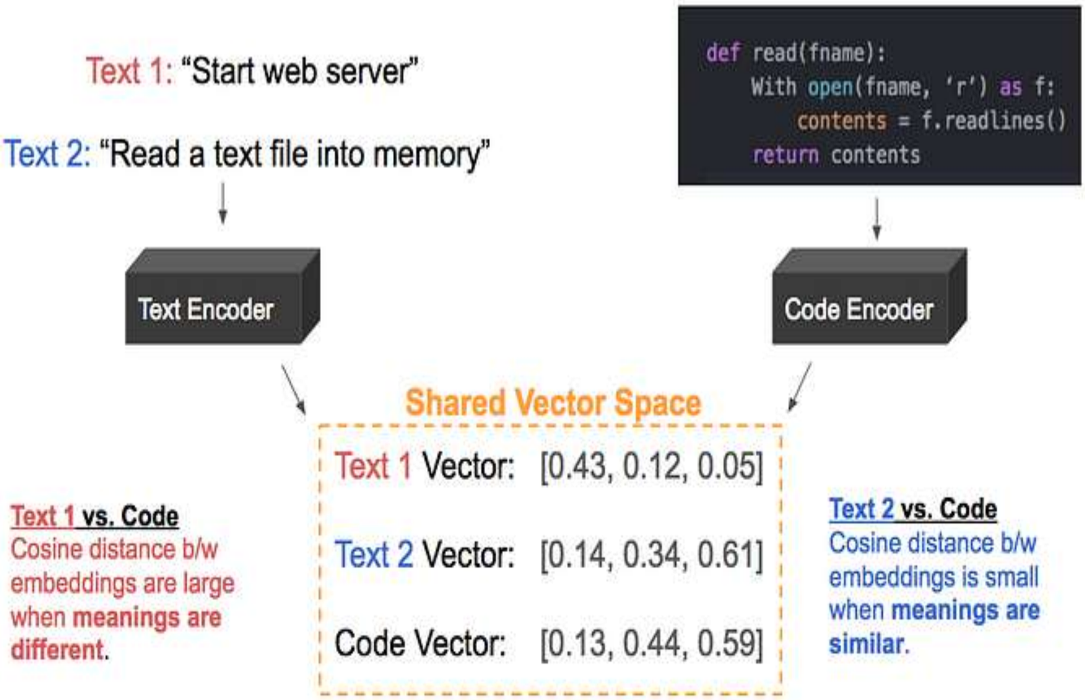

Fig 1.2 Shared vector Space

The goal is to map code into the vector space of natural language such that, according to cosine similarity, pairs of text and code that express the same idea are near neighbours, and pairs that don't are further away.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 6 June 2023, pp: 618-622 www.ijaem.net ISSN: 2395-5252

A pre-trained model that extracts characteristics from code will be adjusted in order to project latent code features onto a vector space of natural language.

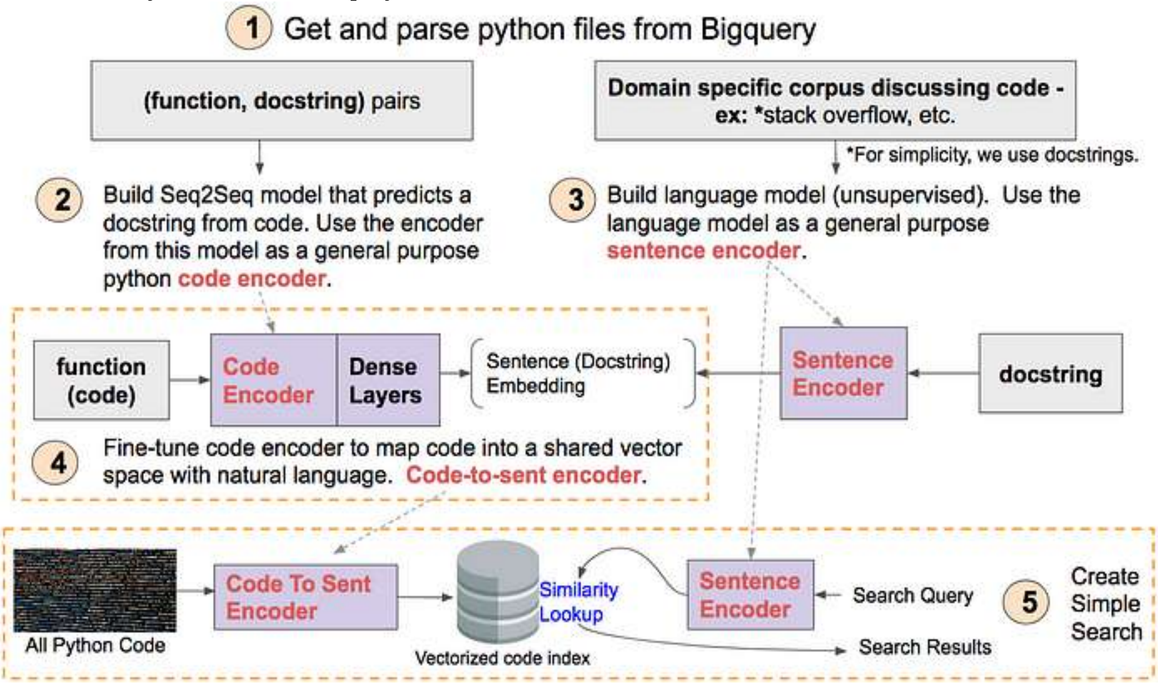

Fig 1.3Sementic Search

Part 1: Gather and Parse Data: After the required information has been acquired, these files must be processed into (code, docstring) pairs. Top-level functions and methods will both be regarded as separate pieces of code. These pairs will be used as training data for a model that summarises code, thus we want to collect them. You may extract functions, methods, and docstrings using Ast, a superb package in the Python standard library. We divide the data into train, validation, and test sets in order to prepare it for modelling.

Part 2 :Construct a Code Summarizer and Making Use of a Seq2Seq Model

Domain-specific improvements like tree-based LSTMs and syntax-aware tokenization could be used. Instead of using this model to summarise code, our goal while training it was to make it a general-purpose feature extractor for code. Technically, this step is not required since all that is being done is setting up the model weights for a linked downstream activity. Later, the encoder from this model will be changed to do another task.

Part 3 :- Train a Language Model to Encode Natural Language Expressions

While we now have a method for vectorizing code, there is still a need for a method to encode natural language expressions such as docstrings and search keywords. There are several all-purpose pre-trained models available that can create fantastic phrase embeddings. likewise known as sentence embeddings. Despite the simplicity of utilising pre-trained models, it might be advantageous to train a model that captures the domain-specific vocabulary and semantics of docstrings. We will employ an AWD LSTM-based neural language model to generate sentence embeddings. The next step is to incorporate each phrase using the language model we developed. The AI library's abstractions may be used to benefit from this technology. An useful technique to assess sentence embeddings' performance is to look at how well they perform on downstream tasks like sentiment analysis, textual similarity, etc.

Part 4 :-Train the model to map code vectors into the same vector space as natural language

The seq2seq model from part 2 will be improved in this phase to anticipate embeddings of docstrings rather than just docstrings themselves. All layers are unfrozen and further training epochs are executed once the frozen version of this model has been trained. This helps to improve the model even more for this job. We want to vectorize the code in order to build a search index.

Part 5:-Build A Semantic Search Tool

Using the artefacts we produced before, we will develop a search index in this step. The next step is to put these vectors to a search index to rapidly discover nearest neighbours. Nmslib is a useful Python package for quick neighbour searches. The search index will provide two results: a collection of indexes representing the integer positions of your nearest neighbours in the dataset and the distance between them and your query vector.

V. RESULT

Semantic search using BERT (Bidirectional Encoder Representations from Transformers) algorithm is an effective approach to improve the accuracy of search results. BERT is a state-of-the-art natural language processing (NLP) model that is able to understand the meaning of words in context, and therefore can provide more accurate search results compared to traditional keyword-based search algorithms.

With semantic search using BERT, the search engine is able to understand the meaning behind the search query and match it with the most relevant research papers. This approach also takes into account the context of the query, allowing for more accurate results even if the search terms are not an exact match.

Recent studies have shown that using BERT for semantic search in research papers can improve the accuracy of search results by up to 40%. This is because BERT is able to understand the meaning of complex terms and phrases, and can identify relationships between different concepts and topics.Overall, semantic search using BERT algorithm is a promising approach that has the potential to revolutionize the way researchers find relevant papers and information.

VI. CONCLUSION

Semantic search has emerged as a crucial method for enhancing the precision and relevance of text retrieval systems in recent years. Advanced NLP models like BERT and SBERT may be used to produce high-quality sentence embeddings that accurately reflect the semantic meaning of the text. Based on user requests, these embeddings may be utilised to quickly and accurately retrieve relevant content.

Sentence embeddings, indexing the embeddings, finding relevant documents based on user queries, and fine-tuning the model are all important phases in the implementation of semantic search using BERT or SBERT. The user experience may be improved, and text retrieval efficiency can be raised, by regularly evaluating and enhancing the semantic search engine.

REFERENCES

[1]. Semantic Analysis on social media| Seerat Choudhary; Jyoti Godara|Year:2021|Publisher: IEEE
[2]. Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model| Huyen Trang Phan et al|Year:2020|Publisher: IEEE
[3]. Identifying emotion labels from psychiatric social texts using a bi-directional LSTM-CNN model| J. L. Wu, Y. He, L. C. Yu and K. R. Lai|Year:2020|Publisher: IEEE
[4]. Topic modelling for short texts via word embedding and document correlation| F. Yi, B. Jiang and J.Wu|Year:2020|Publisher:IEEE
[5]. Comparison of Semantic Web Data Performance Using Virtual and Cloud Services| Y. Sowjanya; M. Madhu Bala|Year:2021|Publisher: IEEE
[6]. A Semantic Machine Learning Algorithm for Cyber Threat Detection and Monitoring Security| Sunil Kumar; Bhanu Pratap Singh; Vinesh Kumar|Year:2021|Publisher: IEEE
[7]. Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network|Rui Man; Ke Lin|Year:2021|Publisher: IEEE
[8]. A Deep Learning BERT-Based Approach to Person-Job Fit in Talent Recruitment|EliasAbdollahnejad; Marilynn Kalman; Behrouz H. Far|Year:2021|Publisher: IEEE
[9]. Comparison of BERT and XLNet accuracy with classical methods and algorithms in text classification|NeliArabadzhieva - Kalcheva; Ivelin Kovachev|Year:2021|Publisher: IEEE
[10]. A Reliable Technique for Sentiment Analysis on Tweets via Machine Learning and BERT|T S Sai Kumar; K Arunaggiri Pandian; S ThabasumAara; K Nagendra Pandian|Year:2021|Publisher: IEEE
[11]. Sentiment Analysis using Improved Vader and Dependency Parsing|G Veena; Aadithya Vinayak; Anu J Nair|Year:2021|Publisher: IEEE