International Journal of Advanced Engineering Science and Information Technology (IJAESIT)
ISSN: 2349-3216
Vol.6, No.6, June 2023

Semantic Search using Transformers

[1]Ashlin Rodrigues
School of Engineering and, Technology
JainUniversity, Bangalore, India
e-mail: accioashlin@gmail.com

[2]Shruti Suresh
School of Engineering and, Technology
JainUniversity, Bangalore, India
e-mail: shrutisureshblr@gmail.com

[3]Akarsh Ghale
School of Engineering and, Technology
JainUniversity, Bangalore, India
e-mail: akarshghale9@gmail.com

[4]Ahmad Alkhuder
School of Engineering and, Technology
JainUniversity, Bangalore, India
e-mail: ahmadalkhuder12@gmail.com

[5]Prof. Venkataravan Nayak K
Department of Computer Science and Engineering
JainUniversity, Bangalore, India
e-mail: nk.venkataravana@jainuniversity.ac.in

[6] Prof. Sunena Rose M V
Department of Computer Science and Engineering
JainUniversity, Bangalore, India
e-mail: sunenarose.m@jainuniversity.ac.in

Abstract— With machines getting increasingly more intuitive with each passing day, there was a demand for search techniques to catch up to a world where technology can consider a variety of scenarios while performing its functions. We shine a light on semantic search, a search technique that can consider the context of the query posed. Its advantages over traditional search techniques are expanded upon, and why there was a need for a shift from those techniques to semantic search. In this project, we have implemented a model that performs semantic search using transformers on research articles that can be provided as document input by the user and presents human-like responses to text queries in a conversational manner, as opposed to chunks of text as in traditional search engines.

Keywords – Semantic search, Transformers, OpenAI, FAISS, Embeddings, NLP

I. INTRODUCTION

In February 1993, students from Stanford University embarked on a project to use statistical analysis of word relationships to improve the efficiency of obtaining search results. In July 1994, a search engine called Lycos worked based on prefix matching and word proximity on, what was at the time, a massive catalogue of indexed documents. In April 1997, a search engine named Ask Jeeves used human editors to provide the most accurate results to queries. Search techniques have come a long way since then, the present iterations owing their efficiency to semantic search. Semantic search refers to the kind of search in which the relationships between words and their contribution to the meaning of the search query is considered. In contrast, traditional search techniques used keyword matching, which returned results according to the number of hits on the keywords entered (the more frequent the keywords in the page, the more relevant).

This sometimes led to search results that were not quite accurate, by failing to understand what the user meant by their query. Semantic search understands the context of the query and returns the most relevant search results accordingly. Of recently, this has been applied to Large Language Models, which allows for users to receive responses to their queries in natural language and negates the need to open several different links to webpages.

II. RELATED WORK

A. Literature Review

The study based on large-scale embedding-based retrieval by Yunkang *et al*.,[1] made it possible for the user to identify relevant information from a large corpus of sentences in a document. They proposed BEBR engine with can binarize and customize bits per dimension avoiding storage and computing issues while dealing with massive documents. The query and embedding are compressed into float vectors using lightweight transformation model. A testing was done to find the effectiveness of the method and it significantly saves upto 50% index cost with almost no loss of accuracy.

According to Zoupanos *et al.*,[2] the objective is to determine which approach is more efficient in comparing sentence embeddings (FAISS and Elasticsearch). Using dataset and measured the performance of the two approaches in terms of efficiency and accuracy.

Zaheer *et al.*,[3] proposed a new type of transformer that is capable of processing longer sequence of text than previous models. The authors propose a new type of transformer called "Big Bird" that uses a combination of global and local attention mechanisms to process longer sequences of text than previous models. They also introduce a sparse attention mechanism that allows Big Bird to scale to much larger input sizes than previous models. They evaluate the performance of Big Bird on a variety of natural language processing tasks, including text classification and question answering, and compare its performance to other state-of-the-art models. The authors also demonstrate that Big Bird is capable of processing much longer sequences of text than previous models, with performance that is competitive with or better than other state-of-the-art models. They also show that Big Bird can be trained efficiently using the sparse attention mechanism, making it a practical tool for processing large amounts of text data. Based on the research paper by Iqbal *et al*.,[4] textual semantic similarity is a very important part when it comes to NLP tasks for example information retrieval, machine translations. This paper was mainly focuses on word embedding techniques to determine the similarities in Bengali sentences. All the word vectors are then made to perform on a dataset containing 50 pairs of Bengali sentences. The results showed that FastText with continuous bag-of-words achieved the highest score of 77.28%.

Introduction of a new pretrained method for transformers that achieves state-of-the art performance on wide range of natural language processing asks was started by Raffel *et al.*,[5] The authors propose a new pretraining method for transformers called the "Text-to-Text Transformer" (T5), which frames all NLP tasks as text-to-text problems. They pretrain T5 on a large corpus of text data using a sequence-to-sequence objective, and then fine-tune it on specific NLP tasks by providing task-specific input and output pairs. They evaluate the performance of T5 on a variety of NLP tasks, including language modeling, question answering, summarization, and translation. To demonstrate that their pretraining method achieves state-of-the-art performance on a wide range of NLP tasks, outperforming previous methods on many tasks. They also show that T5 can be fine-tuned on new tasks with only a small amount of task-specific training data, making it a versatile and effective tool for natural language processing.

Author Beltagy *et al.*,[6] The authors propose a transformer-based model called "Longformer" that is capable of processing long documents by using an attention mechanism that attends to a small number of relevant locations in the input sequence. They also introduce a sparse attention mechanism that enables Longformer to scale to longer input sizes than previous models. They evaluate the performance of Longformer on several natural language processing tasks, including language modeling, question answering, and sentiment analysis, and compare its performance to other state-of-the-art models.

According to Vaswani *et al.*,[7] The objective of the paper is to introduce a new neural network architecture for NLP tasks that uses self-attention mechanisms and does not rely on recurrence or convolutions. Transformer architecture, which uses self-attention mechanisms to weigh the contribution of different input features when making predictions.

The author Masuda *et al.*,[8] Searched documents with integration of NLP techniques to present the effectiveness and flexibility of the framework. The integration of NLP techniques in semantic search can significantly improve the accuracy and effectiveness of the search results.

According to the study made by Wang *et al.*,[9] studied and analyzed the current trends in semantic search and provide a comprehensive view of the field. They investigated several pilot projects and corresponding practical systems.

The survey made by R.Karger *et al*.,[10] was based on the research field of semantic search and provide an overview of the prevalent research directions and common methodologies used in this field Reading and exploring 20 different papers and approaches to semantic search and describe the individual approaches that are part of them, and analyze the common methodologies used in these approaches.

B. Existing Systems

Existing search engines use semantic search, which is a significant upgrade from the older keyword-matching systems. Keyword-matching returned search results that had the most matches with entered search query and ranked by relevance according to the frequency of matches in the returned document.

C. Improvement to Existing Systems

While semantic search has significantly improved the accuracy of search results, users are still required to comb through several different links to acquire the information they seek. Using the proposed model, which utilizes transformers to better judge contextual meaning and hence understand the query better, provides the users with the most relevant information provided in a

concise manner on a single interface, with the information being drawn from throughout the provided document. The response from the model is provided in natural language, rather than being snippets of text from different parts of the document.

III. THE PROPOSED MECHANISM

A. Methodology

Semantic search explains a search engine's trying to generate the most accurate results by understanding depending on the user's intent, context of the query, and the relationship between words. Generative AI is a part of artificial intelligence-based algorithms that is capable to generate new outputs on the data they have been trained on. In this project, we are seeking to implement semantic search on research articles that can be provided as document input by the user, and present human-like responses to the text inputs in a conversational manner, as opposed to returning chunks of text as in traditional search engines. It is based on the most popular and trending application, ChatGPT, which is driven by a large language model (LLM), which means it is programmed/designed in such a way that it can understand human language and generate responses as output based on large corpora of data.

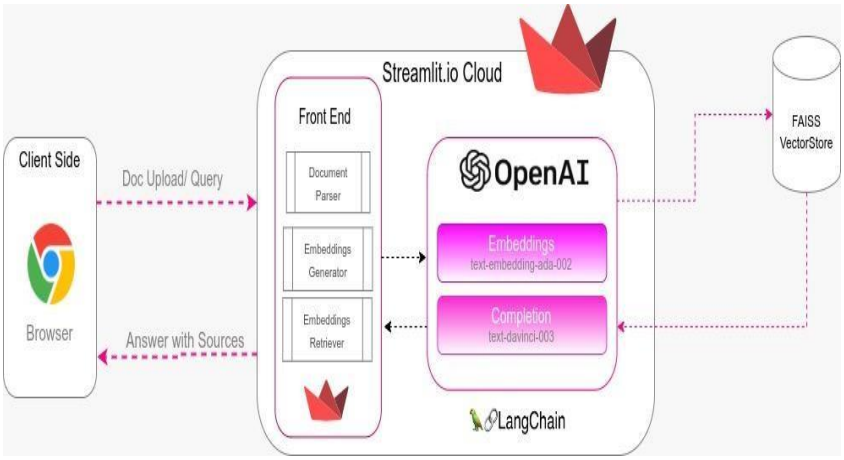


Fig1. Architecture

This architecture stands out as the best choice for several reasons, primarily due to its effective amalgamation of state-of-the-art natural language processing techniques and efficient information retrieval mechanisms. Below, we outline the key advantages of this architecture:

1. *Flexibility*: By accepting user-uploaded documents in popular file formats such as .txt and .pdf, the system caters to a wide range of user needs and content types.

2. *Granular text processing*: The sophisticated parsing algorithm enables the system to handle large documents by breaking them down into smaller, manageable chunks, thereby improving the efficiency of the subsequent processing steps.

3. *Semantic understanding*: Leveraging the text-embeddings-ada-002 model for generating high-dimensional vector embeddings allows the system to capture the semantic meaning of the text chunks, resulting in a deeper understanding of the content and more accurate query responses.

4. *Scalability*: Utilizing the Faiss vector database for storing embeddings, text chunks, and source information ensures high-performance retrieval even as the database grows, making it a suitable choice for large-scale deployments.

5. *Context-aware response generation*: By employing GPT-3.5-Turbo, a cutting-edge language model, the system can synthesize well-informed responses that consider the context provided by the top-K most relevant text chunks, leading to highly accurate and coherent answers.

6. *Source attribution*: The architecture's ability to incorporate source information in the response generation process ensures that the generated responses are not only accurate but also properly attributed, promoting transparency and credibility.

7. *User experience*: The Streamlit-powered front-end interface provides an intuitive and interactive platform for users to submit queries and receive responses, enhancing the overall user experience.

In summary, this architecture capitalizes on the strengths of modern natural language processing and efficient information retrieval techniques to create a sophisticated, professional, and highly effective system for addressing user queries. Its flexibility, scalability, and focus on delivering accurate and well-informed responses make it the best choice for a wide range of applications.

This is a high-level overview of the architecture:

In this paper, we present a unique architecture for an advanced information retrieval and response generation system, organized into the following steps:

1. Accepting user-uploaded documents in popular file formats, such as .txt and .pdf.

2. Utilizing a sophisticated parsing algorithm to extract text content and segment it into discrete chunks.

3. Employing a state-of-the-art transformer-based language model, text-embeddings-ada-002, to generate high-dimensional vector embeddings for the text chunks.

4. Storing the embeddings, accompanied by their respective text chunks and source information, in a performant and scalable vector database, Faiss.

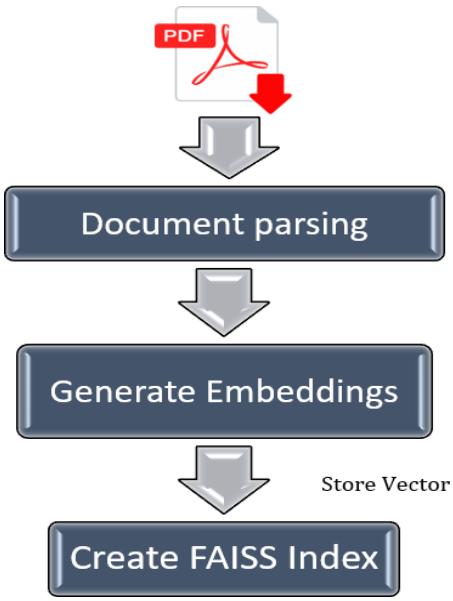


Fig2. Flowchart

Upon user query submission through the Streamlit-powered front-end interface, the system proceeds with the following steps:

5. Applying the same text-embeddings-ada-002 model to obtain a vector representation of the user query.

6. Computing a cosine similarity metric between the query vector and the vectors in the Faiss database to ascertain the top-K most relevant text chunks.

7. Incorporating the pertinent chunks as context into the prompt of the cutting-edge GPT-3.5-Turbo language model, developed by OpenAI.

8. Generating a comprehensive and well-informed response to the user query while maintaining proper citation of sources.

9. Returning the synthesized response to the user via the front-end interface, complete with its corresponding source attributions.

This architecture amalgamates the power of modern natural language processing techniques with efficient information retrieval mechanisms, resulting in a sophisticated and professional system for addressing user queries.

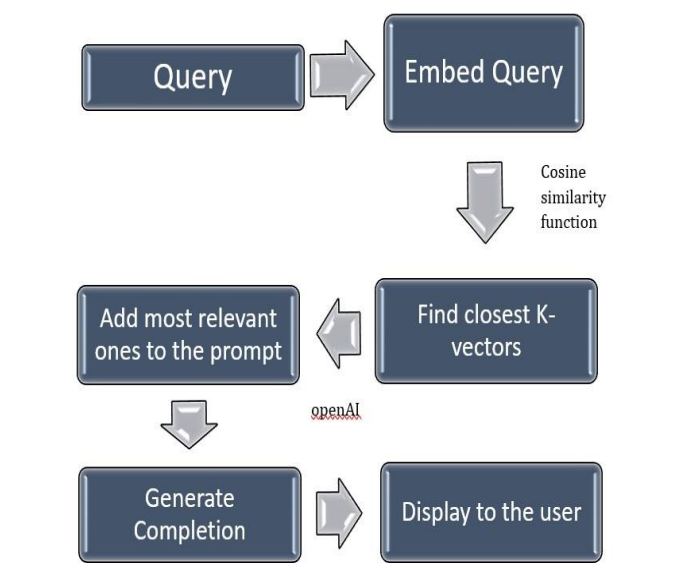


Fig3. Query Flowchart

B. Softwares Used

i. LangChain

LangChain is a framework mainly used for developing applications that are strongly based on language models. They are built around LLMs and uses utility tool library. The main idea of this library is that different components can be "chain" together to create more opportunities of cases using LLMs. LLMs models are rising as transformative technology. This was mainly built to integrate seamlessly with LangChain python package.

ii. Streamlit.io

Streamlit is an open-source python-based framework mainly used for Front-end. It uses python-based library mainly designed for machine learning and data science-based web apps. Since it is built on top of python, it supports mainstream python libraries like pandas, matplotlibs, etc.

iii. OpenAI models

OpenAI models are non-deterministic model, which means that any identical inputs can give different outputs. The API is powered by a set of models with different capabilities. We have used two types.

a. Embeddings – It is used to create embeddings. An embedding model will factorize the input into vector and that vector will be used to predict the next move.

b. GPT 3.5Turbo – This model accepts a set of messages as input. It provides multiple features such as the ability to store prior responses or query with a predefined set of instructions with context.

iv. Faiss

Faiss is a similarity search vector database. It contains algorithms that is capable of searching in groups of vectors of any vector-size and contains code that supports for evaluation and parameter tuning. Usually, we index the vector using Faiss- then using another vector also known as query vector, we search for the most similar vectors within the index.

C. The Working Website

Academic Assistant is made easy for researchers by allowing the user to ask question about any document and get back accurate results with instant citation.

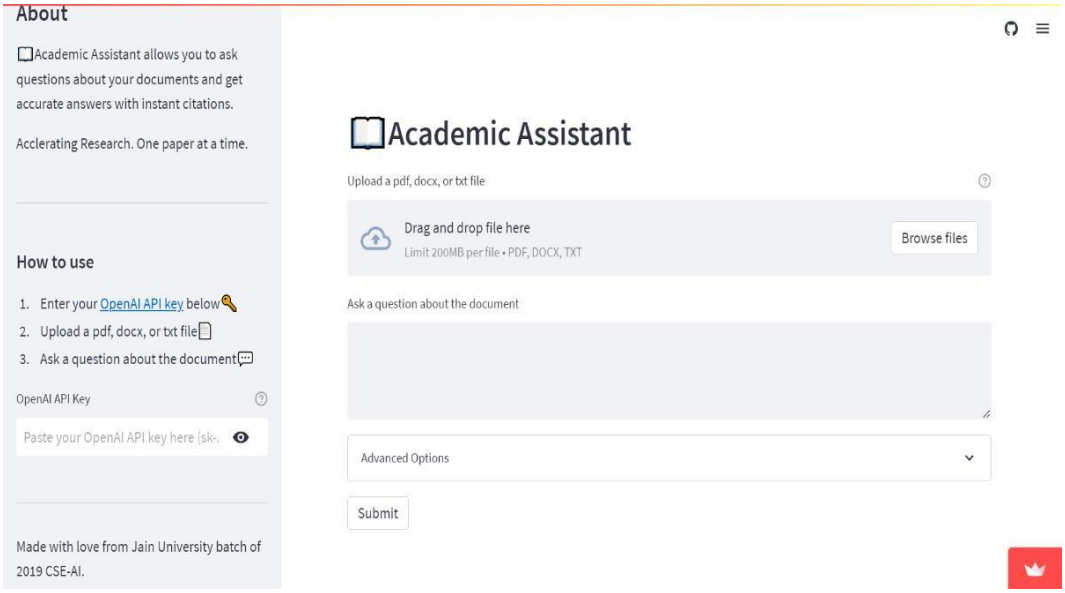


Fig 4a. Academic Assistant

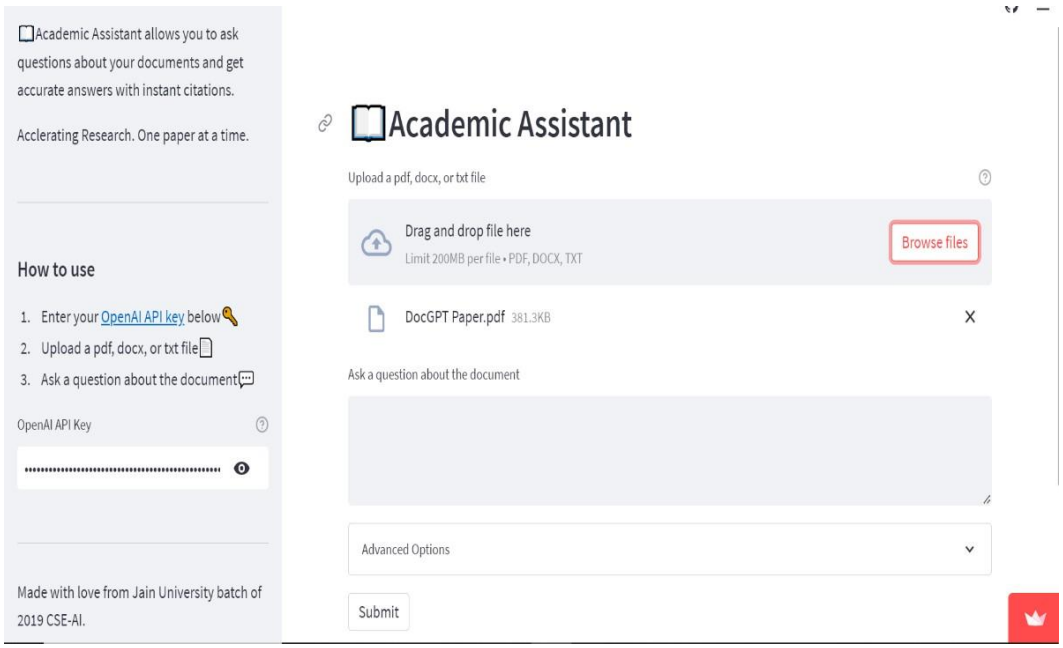


Fig 4b. Generate API key and Upload a document

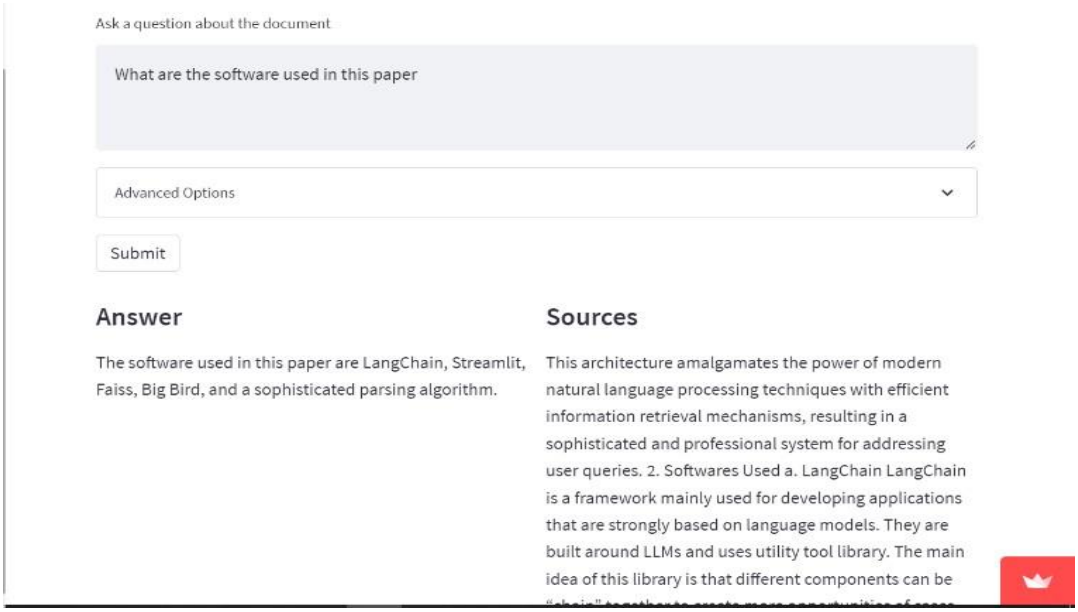


Fig 4c. Ask a question and answer is extracted from the document.

IV. PERFORMANCE EVALUATION

In this section, we analyze the results obtained by comparing two Search techniques, implemented using Google Colab.

The two search techniques are:

1. Keyword Search technique
2. Semantic Search technique

We will be evaluating the accuracies of both the search techniques without and with transformers, using distilbert-base-uncased for generating embeddings. The dataset used in this experiment is Quora question pairs dataset that contains 10K samples. The goal of this comparison is to predict which of the provided pairs of questions contain two questions with the same meaning.

Performance metric that was considered is:

1. Accuracy

Accuracy is the ratio of number of correct predictions to the total number of input samples. It is a metric that generally describes how the model performs across all classes.

$$\text{Accuracy} = \frac{\text{True}_{\text{positive}} + \text{True}_{\text{negative}}}{\text{True}_{\text{positive}} + \text{True}_{\text{negative}} + \text{False}_{\text{positive}} + \text{False}_{\text{negative}}}$$

The comparisons presented in tables and graphs below.

Search technique/ Metric	***Accuracy***
Keyword Search	55%
Semantic Search	60%

Table 1. Comparison of search techniques using Static Word Embedding

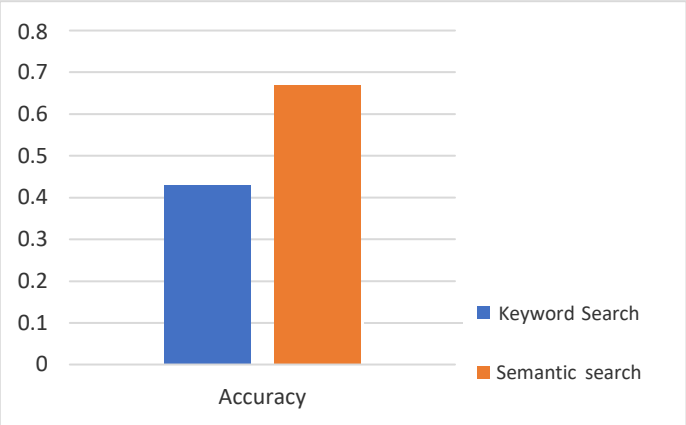


Fig 5. Comparison of search techniques using Static Word Embedding

Search technique/ Metric	***Accuracy***
Keyword Search	55%
Transformer-based Semantic Search	91%

Table 2. Comparison of search techniques using Dynamic Word Embedding

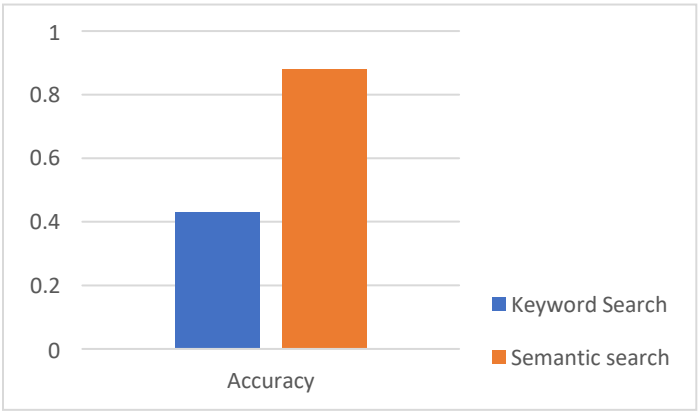


Fig 6. Comparison of search techniques using Dynamic Word Embedding

- Static Word Embeddings

Examples of models that result in static word embeddings include Word2Vec, GloVe, FastText, etc. Static Word Embeddings fail to capture polysemy. They generate the same embedding for the same word in different contexts. They only incorporate previous knowledge in the first layer of the model.

- Dynamic Word Embeddings

Examples of models that result in dynamic word embeddings include BERT, CoVe, GPT, etc. Contextualized embeddings such as BERT have been shown more effective than static embeddings as NLP input embeddings. Such embeddings are it dynamic, calculated according to a sentential context using a network structure.

From the above tables and graphs, we infer that transformers improves the performance of these search techniques. The best performing search technique is the Semantic search, whose accuracy have improved with dynamic word embeddings, from 67% to 88%

V. APPLICATION

Using semantic search, we have created a model that searches more intuitively, while also presenting results in natural language. It can be used as an intelligent document assistant, suitable for a variety of applications.

1. Students and academics can provide research articles as input and use queries to extract the relevant information from them.
2. Companies and corporations, like law firms, often have huge amounts of documentation to handle. The information required can be fetched by the model from those documents.
3. Medical reports can be parsed and queried to receive the required information from them.

VI. FUTURE SCOPE

This model can be used to extract specific pieces of information without having to read the entire document. This can be extended to documents other than research articles too. In the future, the model could be amended to model could be amended to handle more than one document at a time, and fetch results considering the contents of both documents.

VII. CONCLUSION

The best performing search technique is the Transformer-based Semantic search, whose accuracy have improved with dynamic word embeddings, from 60% to 91%

Using semantic search, we have created a model that searches more intuitively, while also presenting results in natural language. It can be used as an intelligent document assistant, suitable for a variety of applications.

ACKNOWLEDGEMENT

The authors are grateful for the constant support and guidance of M r . Venkataravana Nayak k and Mrs.Sunena Rose throughout the implementation of the project. The authors would also like to thank School of Engineering and Technology, Jain University, Bangalore for its massive infrastructure and access to all the amenities needed in the completion of this project.

REFERENCES

[1] Gan, Yukang, Yixiao Ge, Chang Zhou, Shupeng Su, Zhouchuan Xu, Xuyuan Xu, Quanchao Hui, Xiang Chen, Yexin Wang, and Ying Shan. "Binary Embedding-based Retrieval at Tencent." arXiv preprint arXiv:2302.08714 (2023).

[2] Zoupanos, Spyros & Kolovos, Stratis & Kanavos, Athanasios & Papadimitriou, Orestis & Maragoudakis, Manolis. (2022). Efficient comparison of sentence embeddings.

[3] Zaheer, Manzil, et al. "Big bird: Transformers for longer sequences." Advances in neural information processing systems 33 (2020): 17283-17297.

[4] MD. Asif Iqbal, Omar Sharif, Mohammed Moshiul Hoque, Iqbal H. Sarker, Word Embedding based Textual Semantic Similarity Measure in Bengali, Procedia Computer Science, Volume 193, 2021, Pages 92-101, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.10.010.

[5] "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.

[6] Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[8] Katsuya Masuda, Takuya Matsuzaki, Jun'ichi Tsujii,Semantic Search based on the Online Integration of NLP Techniques,Procedia - Social and Behavioral Sciences,Volume 27,2011,Pages 281-290,ISSN 1877-0428

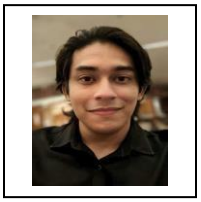
[9] Wei, Wang & Barnaghi, Payam & Bargiela, Andrzej. (2008). Search with meanings: An overview of semantic search systems. International Journal of Communications of SIWN. 3.

[10] Mäkelä, Eetu. (2008). Survey of Semantic Search Research.

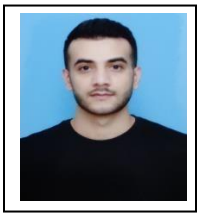
Author's Profile

Ashlin Rodrigues is an undergraduate student currently pursuing a degree in Computer Science and Engineering, specializing in Artificial Intelligence from Jain University, Bangalore. Her areas of interests are Artificial Intelligence and Machine Learning. She has ventured into projects on collaborative softwares. She is currently doing an internship at a global pharmaceutical company.

Shruti Suresh is an undergraduate student currently pursuing a degree in Computer Science and Engineering, specializing in Artificial Intelligence from Jain University, Bangalore. Her interests included data analytics and cloud technologies. She has internship experience at a global healthcare company.



Akarsh Ghale is a Computer Science undergraduate studying at SET, Jain University, Bangalore. His specialization is in Artificial Intelligence, and interests include programming and research.



Ahmad Alkhuder is a Computer Science undergraduate studying at SET, Jain University, Bangalore. His specialization is in Artificial Intelligence and pursue master's degree in the same field.

Mr. Venkataravana Nayak K received B.E. degree and M.E. degree in computer science and engineering from Visvesvaraya Technological University and Bangalore University in 2003 and 2006 respectively. He has worked towards Ph.D. degree in computer science and engineering, Bangalore University. He has worked in various engineering colleges and currently working as Assistant Professor in the Department of Computer Science and Engineering, School of Engineering and Technology, Jain University, Bangalore, India. He has published more than 25 papers in journals and conferences of international standard, published two books and three patents. The research interests include Digital Image Processing and Artificial Intelligence.

Mrs. Sunena Rose M V received M.Tech degree in Computer Science and Engineering. She is currently working as a Assistant Professor (CSE-AI) In Jain University and also Pursuing PhD. She has 4 years of experience in Research & Academics. Area of Interest-Artificial Intelligence and Machine Learning.