EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Enhancing User Query Comprehension and Contextual Relevance with a Semantic Search Engine using BERT and ElasticSearch

Saniya M Ladanavar[1], Ritu Kamble[1], R.H Goudar[1,*], Rohit.B. Kaliwal[1], Vijayalaxmi Rathod[1], Santhosh L Deshpande[1], Dhananjaya G M[1] and Anjanabhargavi Kulkarni[1]

[1]Dept. of CSE, Visvesvaraya Technological University, Belagavi, Karnataka, India

Abstract

This research paper explores the development of a semantic search engine designed to enhance user query comprehension and deliver contextually applicable research results. Classic search engines basically struggle to catch the nuanced meaning of user queries, giving to suboptimal results. To address this challenge, we give the merge of advanced natural language processing (NLP) techniques with Elasticsearch, and with a specific focus on Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art pre-trained language model. Our approach involves leveraging BERT's ability to analyze the contextual meaning of words within documents by sentence transformers as (SBERT) , enabling the search engine to grab the user queries and better under- stand semantics of the content as it is converted into vector embeddings making it understandable in the Elasticsearch server. By utilizing BERT's bidirectional attention mechanism, the search engine can discern the relationships between words, thereby capturing the contextual nuances that are crucial for accurate query interpretation. Through experimental validation and performance assessments, we demonstrate the efficacy of our semantic search engine in providing contextually relevant search results. This research contributes to the advancement of search technology by enhancing the intelligence of search engines, ultimately improving the user experience by giving context based research.

Keywords: Semantic Search Engine, Elastic Search , BERT Model (SBERT)

Received on 9 April 2024, accepted on 24 June 2024, published on 21 August 2024

doi: 10.4108/eetiot.6993

1. Introduction

Semantic search engines helps in understanding the context and intent behind user queries, giving more accurate and context-aware search results. This paper presents an investigation into the integration of Elasticsearch, a widely-used search engine, with BERT, a state-of-the-art language model, to build a semantic search engine.

The motivation behind constructing a semantic search engine using Elasticsearch and the BERT model mbeddings from the imperative to elevate the precision and contextual relevance of search results. Conventional search engines often grapple with the limitation of relying solely on keyword matching, leading to suboptimal outcomes when users' queries are nuanced or context-dependent. Elasticsearch, renowned for its robust full-text search capabilities and scalability, provides a robust foundation for efficiently indexing and retrieving documents. The integration of the BERT model, known for its prowess in comprehending the intricacies of language and contextual nuances, aims to revolutionize the search paradigm. By incorporating BERT into the search process, the engine can discern the semantic meaning of words in a sentence, thereby facilitating a more profound understanding of user intent. This amalgamation not only addresses the shortcomings of traditional search methods but also promises to deliver search results that correspond with the user's intended meaning, ushering in a new era of semantic search that enhances the overall search experience, efficiency, and user satisfaction. The existing landscape of search engines

*Corresponding Author. Email: rhgoudar.vtu@gmail.com

The significance of semantic search engines lies in their capacity to revolutionize the search experience by transcending the limitations of traditional keyword-based systems. Semantic search engines, powered by advanced technologies like Elasticsearch and the BERT model, prioritize context and meaning, allowing for a more refined understanding of the queries by users. This enhanced comprehension enables the delivery of results that go beyond mere keyword matching, addressing the intricacies of language, synonyms, and the enlarged context of the search terms. Importantly, semantic search engines contribute to a bigger personalized and user- centric online experience. By discerning the intention behind query, these engines can tailor results to align precisely with user expectations, ultimately not wasting time and giving more personalized information In an era of information abundance, where precise and context-aware search is crucial, semantic search engines emerge as indispensable tools, offering a transformative approach to information retrieval that significantly improves user satisfaction and engagement. Their importance extends across diverse domains, from set on a separate line.

2. Literature Overview

The paper [1] highlights the limitations of keyword-based search engines and advocates for the virtue of Semantic Web and Search Engines. It introduces SemanTelli, a meta- semantic search engine utilizing intelligent agents to aggregate results from various semantic search engines. Proposed enhancements for SemanTelli include an improved snippet analysis-based page ranking algorithm and the addition of image and news search functionalities. These upgrades aim to increasing the precision and comprehensiveness of results, positioning SemanTelli as an advanced solution in web search. The paper [2] proposes an innovative solution for academic search by aggregating results from diverse sources using BERT contextual embedding's. This approach, a first in academic search, significantly improves retrieval performance. Experimental results highlight the superiority of BERT over other language models, emphasizing its advantage in optimizing academic aggregated search systems. The paper [3] proposes an innovative solution for academic search by aggregating results from various sources, utilizing unsupervised BERT contextual embedding's. This pioneering approach significantly improves retrieval performance, outperforming other language models such as ELMo, USE, and XLNet. The findings highlight advantages of the BERT language model in optimizing academic aggregated search systems.. The paper [4] highlights the limitations of existing search engines in handling large and complete queries. The author introduces a semantics-oriented search engine utilizing neural networks and BERT embedding's, demonstrating improved accuracy in ranking documents based on query relevance. Key terms include Deep Neural Networks (DNN), Bidirectional Encoder Representations from Transformer (BERT), cosine similarity, Long-Short Term Memory (LSTM), and Siamese LSTM, showcasing the incorporation of advanced techniques for enhanced search performance. The paper [5] introduces an intelligent search method for energy enterprises, utilizing the Bert preprocessing model on heterogeneous data. The approach combines template matching and text classification for intention recognition, bridging machine learning and deep learning within artificial intelligence. The preprocessing model, Bert, transforms natural language into a vector based on syntax, while information extraction technology extracts structured information for intention processing parameters. The intention recognition employs both template matching and text classification methods. Experimental examples are used to compare the effectiveness of these approaches. The paper [6] addresses the challenges in document retrieval for search engines, particularly when optimizing with verbose or tail queries. It proposes a vector space search framework utilizing a deep semantic matching model trained on the BERT architecture. The model encodes each query and document into a low dimensional embedding. The implementation includes a fast nearest neighbor index service for efficient online serving. Both offline and online metrics indicate a significant improvement in retrieval performance and search quality, particularly for tail queries. The paper [7] presents a personalized news search engine utilizing efficient text extraction from web news pages. Employing DOM tree manipulation, the system removes irrelevant content, such as ads and comments. Semantic matching with Word- Net enhances content relevance, and TF-IDF identifies information-carrying blocks. The system gathers related information from diverse web news sources and provides summarization based on user preferences. Results show good recall and high precision for both generalized and specific queries. The paper [8] discusses the limitations of current search engines in meeting the demands of informatics search in knowledge intensive fields like medicine and biology. The solution proposed is TCM Search, a semantic based search engine designed for traditional medical informatics. TCM Search utilizes Semantic Web techniques to enhance the informatics search experience in various aspects, addressing the challenges in knowledge-intensive disciplines before achieving breakthroughs in AI or NLP. The paper [9] introduces the concept of Question Answering (QA) systems in information retrieval, highlighting their task of automatically providing correct answers to natural language questions

The paper [10] discusses various methodologies and implementation details for a general language QA system. Uses more required answers using Natural Language Processing (NLP) techniques. The text addresses the challenge of navigating vast biomedical information, citing the example of PubMed with over 27 million publications. While PubMed uses a keyword-based search, it presents results as a paginated textual list, which can be time-consuming to navigate. To enhance user experience, various interfaces, including ViLiP, have emerged. ViLiP, developed as a visual exploratory interface for PubMed, is primarily used in neuroscience. It presents query results as an inside heatmap. This work extends ViLiP by incorporating an NLP- based semantic search engine, specifically focusing on detecting drug information within queries. The paper [11] explores an intelligent search method related on a knowledge graph, encompassing the construction of the knowledge graph, strategic word segmentation, retrieval keyword input, keyword processing, and outputting search results. It employs NLP semantic understanding and ElasticSearch (ES) for search, establishing a modularized retrieval data model. The data model is organized depending on content, data type, and data characteristics, improving the internal data structure. This modular approach enhances the search speed, particularly for handling vast amounts of data information

3. Methodology

The architecture follows a well-defined process for constructing an effective search engine. Elasticsearch serves as the core infrastructure, organized with an intricately de- signed index accommodating CSV data. The inclusion of Pandas facilitates meticulous data preprocessing, ensuring data integrity and relevance. Integration with the SBERT model adds a layer of semantic understanding, converting textual information into vector embedding's for enhanced search capabilities. Mapping the index structures ensures seamless compatibility with Elasticsearch settings, optimizing the efficiency of data storage and retrieval. The Streamlit UI frontend, developed for intuitive interaction, provides users with a straightforward interface to query and visualize search results. Rigorous testing and performance optimization contribute to a robust and reliable search engine. The required considerations are embedded throughout the process, prioritizing privacy and consent. The documentation, detailing configurations, preprocessing steps, and UI development, serves as a valuable resource for future maintenance and improvements. Overall, this architecture embodies a systematic and comprehensive approach to constructing a sophisticated search engine with a focus on accuracy, efficiency, and user experience.

The methodology employed in this project serves as a comprehensive plan outlining the principles, processes, and rules utilized during the development of an Elasticsearch based search engine. This delineates the step-by-step approach to the entire process, detailing the setup of Elasticsearch as the core infrastructure, the creation and mapping of indices to organize and store CSV data, and the integration of Pandas for meticulous data preprocessing. It further encompasses the incorporation of SBERT model embed- dings to enhance semantic understanding and the definition of mapping structures to ensure compatibility with Elasticsearch settings. Additionally, the methodology out- lines the systematic development of a Streamlit UI frontend for an intuitive user experience. In essence, the methodology provides a structured "how-to" guide for conducting each phase of the project, ensuring clarity, repeatability, and the achievement of specific research or project objectives.

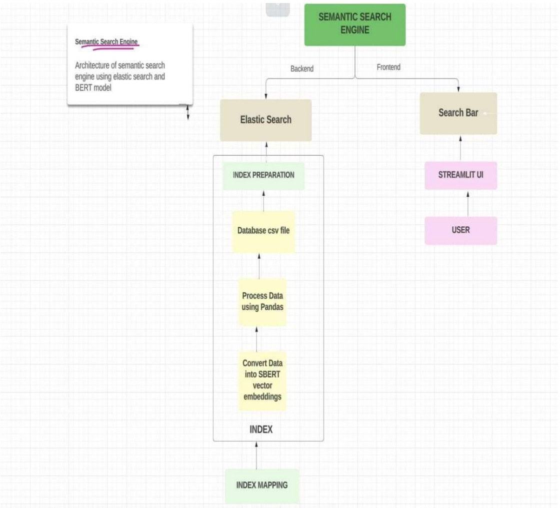


Figure 1. The architecture follows a well-defined process for semantic search engine

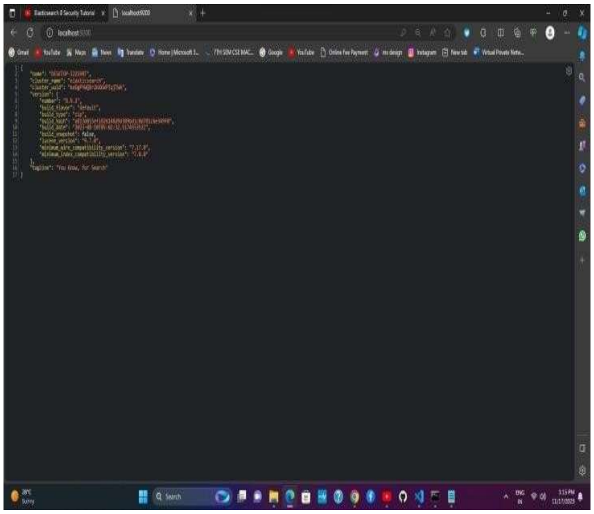
This project involves a meticulous sequence of steps, strategically chosen to achieve the research objectives of creating an efficient Elasticsearch based search engine. The process began with the setup of Elasticsearch, where the installation and configuration laid the groundwork for a robust infrastructure.

Subsequently, an Elasticsearch index was meticulously prepared to organize and store the dataset, and its mapping structures were defined to ensure compatibility with Elasticsearch settings. Data process of database csv file was conducted using Pandas, importing CSV data into a DataFrame and applying preprocessing techniques to uphold data integrity. The integration of SBERT model embedding's was a crucial step, enhancing the search engine's capabilities with semantic understanding. Index mapping was then implemented, specifying data types and structures for optimized data retrieval. The actual data insertion into Elasticsearch and the development of a Streamlit UI frontend followed suit, providing an interactive platform for users to query and visualize search results seamlessly. Rigorous testing, performance optimization, and thorough documentation were integral components of the overall strategy, ensuring the reliability, efficiency, and maintainability of the search engine. This comprehensive methodology not only details the tools and steps taken but also reflects a systematic and thoughtful approach to achieving the research objectives.

4. Result

4.1.ElasticSearch setup

Installed and configured Elasticsearch to serve as the search engine infrastructure and established the necessary connections and settings for optimal performance as shown in the Figure 2.

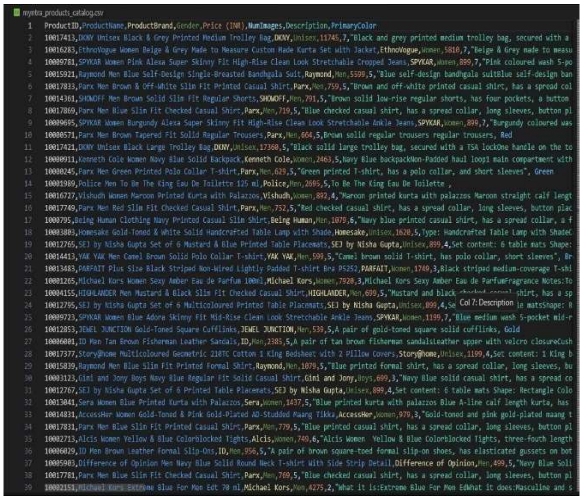


```
{
    "name": "[illegible]",
    "cluster_name": "elasticsearch",
    "cluster_uuid": "[illegible]",
    "version": {
        "number": "[illegible]",
        "build_flavor": "default",
        "build_type": "zip",
        "build_hash": "[illegible]",
        "build_date": "[illegible]",
        "build_snapshot": false,
        "lucene_version": "[illegible]",
        "minimum_wire_compatibility_version": "7.17.0",
        "minimum_index_compatibility_version": "7.0.0"
    },
    "tagline": "You Know, for Search"
}
```

Figure 2. Represents the installation and hosted Elasticsearch server

4.2. Index Preparation

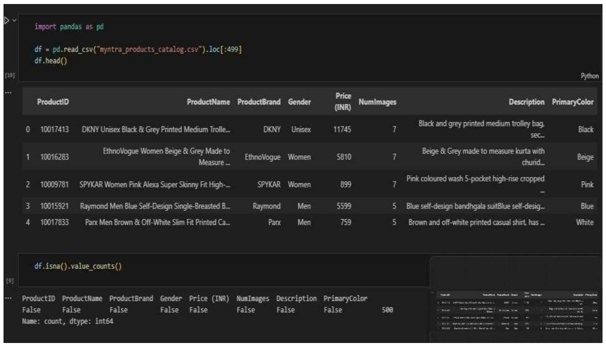
Created an index in Elasticsearch to organize and store the dataset and mapped the index structure to accommodate the data from the CSV file as shown in the Figure 3.



```
ProductID,ProductName,ProductBrand,Gender,Price (INR),NumImages,Description,PrimaryColor
10017413,DKNY Unisex Black & Grey Printed Medium Trolley Bag,DKNY,Unisex,11745,7,"Black and grey printed medium trolley bag, secured with a
10016283,EthnoVogue Women Beige & Grey Made to Measure Custom Made Kurta Set with Jacket,EthnoVogue,Women,5810,7,"Beige & Grey made to measu
10009781,SPYKAR Women Pink Alexa Super Skinny Fit High-Rise Clean Look Stretchable Cropped Jeans,SPYKAR,Women,899,7,"Pink coloured wash 5-po
10015921,Raymond Men Blue Self-Design Single-Breasted Bandhgala Suit,Raymond,Men,5599,5,"Blue self-design bandhgala suitBlue self-design ban
10017833,Parx Men Brown & Off-White Slim Fit Printed Casual Shirt,Parx,Men,759,5,"Brown and off-white printed casual shirt, has a spread col
10014361,SHOWOFF Men Brown Solid Slim Fit Regular Shorts,SHOWOFF,Men,791,5,"Brown solid low-rise regular shorts, has four pockets, a button
10017869,Parx Men Blue Slim Fit Checked Casual Shirt,Parx,Men,719,5,"Blue checked casual shirt, has a spread collar, long sleeves, button pl
10009695,SPYKAR Women Burgundy Alexa Super Skinny Fit High-Rise Clean Look Stretchable Ankle Jeans,SPYKAR,Women,899,7,"Burgundy coloured was
10000571,Parx Men Brown Tapered Fit Solid Regular Trousers,Parx,Men,664,5,Brown solid regular trousers regular trousers, Red
10017421,DKNY Unisex Black Large Trolley Bag,DKNY,Unisex,17360,5,"Black solid large trolley bag, secured with a TSA lockOne handle on the to
10000911,Kenneth Cole Women Navy Blue Solid Backpack,Kenneth Cole,Women,2463,5,Navy Blue backpackNon-Padded haul loop1 main compartment with
10000245,Parx Men Green Printed Polo Collar T-shirt,Parx,Men,629,5,"Green printed T-shirt, has a polo collar, and short sleeves", Green
10001989,Police Men To Be The King Eau De Toilette 125 ml,Police,Men,2695,5,To Be The King Eau De Toilette ,
10016727,Vishudh Women Maroon Printed Kurta with Palazzos,Vishudh,Women,892,4,"Maroon printed kurta with palazzos Maroon straight calf lengt
10017749,Parx Men Red Slim Fit Checked Casual Shirt,Parx,Men,752,5,"Red checked casual shirt, has a spread collar, long sleeves, button plac
1000795,Being Human Clothing Navy Printed Casual Slim Shirt,Being Human,Men,1079,6,"Navy blue printed casual shirt, has a spread collar, a f
10003803,Homesake Gold-Toned & White Solid Handcrafted Table Lamp with Shade,Homesake,Unisex,1620,5,Type: Handcrafted Table Lamp with ShadeC
10012765,SEJ by Nisha Gupta Set of 6 Mustard & Blue Printed Table Placemats,SEJ by Nisha Gupta,Unisex,899,4,Set content: 6 table mats Shape:
10014413,YAK YAK Men Camel Brown Solid Polo Collar T-shirt,YAK YAK,Men,599,5,"Camel brown solid T-shirt, has polo collar, short sleeves", Br
10013483,PARFAIT Plus Size Black Striped Non-Wired Lightly Padded T-shirt Bra P5252,PARFAIT,Women,1749,3,Black striped medium-coverage T-shi
10001265,Michael Kors Women Sexy Amber Eau de Parfum 100ml,Michael Kors,Women,7920,3,Michael Kors Sexy Amber Eau de ParfumFragrance Notes:To
10004155,HIGHLANDER Men Mustard & Black Slim Fit Checked Casual Shirt,HIGHLANDER,Men,699,5,"Mustard and black checked casual shirt, has a sp
10012795,SEJ by Nisha Gupta Set of 6 Multicoloured Printed Table Placemats,SEJ by Nisha Gupta,Unisex,899,4,Se                le matsShape: R
10009723,SPYKAR Women Blue Adora Skinny Fit Mid-Rise Clean Look Stretchable Ankle Jeans,SPYKAR,Women,1199,7,"Blue medium wash 5-pocket mid-r
10012853,JEWEL JUNCTION Gold-Toned Square Cufflinks,JEWEL JUNCTION,Men,539,5,A pair of gold-toned square solid cufflinks, Gold
10006001,ID Men Tan Brown Fisherman Leather Sandals,ID,Men,2385,5,A pair of tan brown fisherman sandalsLeather upper with velcro closureCush
10017377,Story@home Multicoloured Geometric 210TC Cotton 1 King Bedsheet with 2 Pillow Covers,Story@home,Unisex,1199,4,Set content: 1 King b
10015839,Raymond Men Blue Slim Fit Printed Formal Shirt,Raymond,Men,1079,5,"Blue printed formal shirt, has a spread collar, long sleeves, bu
10003123,Gini and Jony Boys Navy Blue Regular Fit Solid Casual Shirt,Gini and Jony,Boys,699,3,"Navy Blue solid casual shirt, has a spread co
10012767,SEJ by Nisha Gupta Set of 6 Printed Table Placemats,SEJ by Nisha Gupta,Unisex,899,4,Set content: 6 table mats Shape: Rectangle Colo
10013041,Sera Women Blue Printed Kurta with Palazzos,Sera,Women,1437,5,"Blue printed kurta with palazzos Blue A-line calf length kurta, has
10014831,AccessHer Women Gold-Toned & Pink Gold-Plated AD-Studded Maang Tikka,AccessHer,Women,979,3,"Gold-toned and pink gold-plated maang t
10017831,Parx Men Blue Slim Fit Printed Casual Shirt,Parx,Men,779,5,"Blue printed casual shirt, has a spread collar, long sleeves, button pl
10002713,Alcis Women Yellow & Blue Colorblocked Tights,Alcis,Women,749,6,"Alcis Women  Yellow & Blue Colorblocked Tights, three-fouth length
10006029,ID Men Brown Leather Formal Slip-Ons,ID,Men,956,5,"A pair of brown square-toed formal slip-on shoes, has elasticated gussets on bot
10005903,Difference of Opinion Men Navy Blue Solid Round Neck T-shirt With Side Strip Detail,Difference of Opinion,Men,499,5,"Navy Blue Soli
10017781,Parx Men Blue Slim Fit Checked Casual Shirt,Parx,Men,769,5,"Blue checked casual shirt, has a spread collar, long sleeves, button pl
10002151,Michael Kors Extreme Blue For Men Edt 70 ml,Michael Kors,Men,4275,2,"What it is:Extreme Blue For Men EdWhat it does:Masculine and s
```

Figure 3. Organized the dataset into Elasticsearch accommodate the data

4.3.Data Processing with panda



```python
import pandas as pd

df = pd.read_csv("myntra_products_catalog.csv").loc[:499]
df.head()
```

	ProductID	ProductName	ProductBrand	Gender	Price (INR)	NumImages	Description	PrimaryColor
0	10017413	DKNY Unisex Black & Grey Printed Medium Trolle...	DKNY	Unisex	11745	7	Black and grey printed medium trolley bag, sec...	Black
1	10016283	EthnoVogue Women Beige & Grey Made to Measure ...	EthnoVogue	Women	5810	7	Beige & Grey made to measure kurta with churid...	Beige
2	10009781	SPYKAR Women Pink Alexa Super Skinny Fit High-...	SPYKAR	Women	899	7	Pink coloured wash 5-pocket high-rise cropped ...	Pink
3	10015921	Raymond Men Blue Self-Design Single-Breasted B...	Raymond	Men	5599	5	Blue self-design bandhgala suitBlue self-desig...	Blue
4	10017833	Parx Men Brown & Off-White Slim Fit Printed Ca...	Parx	Men	759	5	Brown and off-white printed casual shirt, has ...	White

```python
df.isna().value_counts()
```

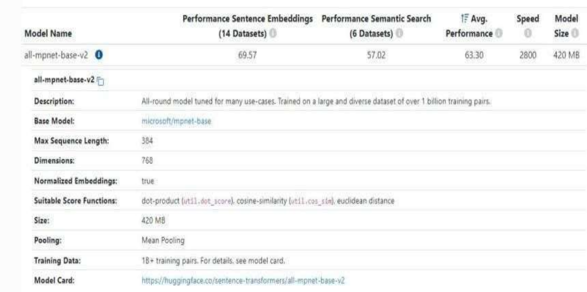
```
ProductID  ProductName  ProductBrand  Gender  Price (INR)  NumImages  Description  PrimaryColor
False      False        False         False   False        False      False        False          500
Name: count, dtype: int64
```

Figure 4. Imported Dataset into Pandas DataFrame

Imported the dataset in CSV format into a Pandas DataFrame and conducted data preprocessing using Pandas to ensure quality and relevance as in Figure 4.

4.4.BERT Model Integration(SBERT)

Utilized SBERT (Sentence-BERT) for generating vector embeddings from relevant data and converted textual information into BERT model embeddings to capture semantic relationships as shown in the Figure 5.



Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
all-mpnet-base-v2	69.57	57.02	63.30	2800	420 MB

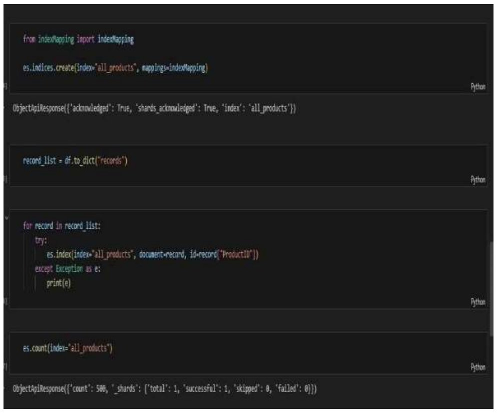
all-mpnet-base-v2

Description:	All-round model tuned for many use-cases. Trained on a large and diverse dataset of over 1 billion training pairs.
Base Model:	microsoft/mpnet-base
Max Sequence Length:	384
Dimensions:	768
Normalized Embeddings:	true
Suitable Score Functions:	dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
Size:	420 MB
Pooling:	Mean Pooling
Training Data:	1B+ training pairs. For details, see model card.
Model Card:	https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Figure 5. Imported Dataset into Pandas DataFrame

4.5.Index Mapping

Defined the mapping for the Elasticsearch index, specifying the data types and structures and ensured compatibility between the indexed data and the chosen Elasticsearch settings as shown in the Figure 6.



```python
from indexMapping import indexMapping

es.indices.create(index="all_products", mappings=indexMapping)
```

```
ObjectApiResponse({'acknowledged': True, 'shards_acknowledged': True, 'index': 'all_products'})
```

```python
record_list = df.to_dict("records")
```

```python
for record in record_list:
    try:
        es.index(index="all_products", document=record, id=record["ProductID"])
    except Exception as e:
        print(e)
```

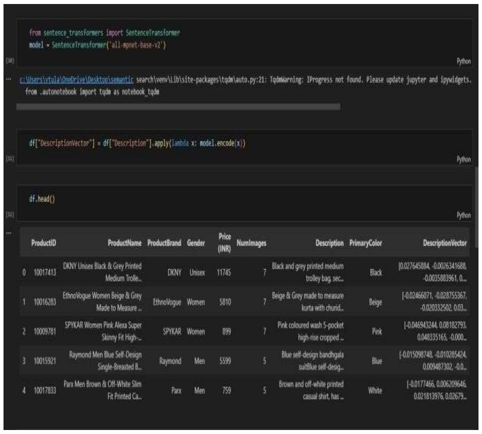
```python
es.count(index="all_products")
```

```
ObjectApiResponse({'count': 500, '_shards': {'total': 1, 'successful': 1, 'skipped': 0, 'failed': 0}})
```

Figure 6. Indexed the Mapping about the dataset in the index

4.6. Indexing Data into Vector embedding's using SBERT

Inserted the preprocessed data into the Elasticsearch index and implemented the description into vector embedding's through SBERT by sentence transformers for better understanding of the context when put as vector embedding's as shown in the Figure 7.



```python
from sentence_transformers import SentenceTransformer
model = SentenceTransformer('all-mpnet-base-v2')
```

```
c:\Users\vtula\OneDrive\Desktop\semantic search\venv\Lib\site-packages\tqdm\auto.py:21: TqdmWarning: IProgress not found. Please update jupyter and ipywidgets.
  from .autonotebook import tqdm as notebook_tqdm
```

```python
df["DescriptionVector"] = df["Description"].apply(lambda x: model.encode(x))
```

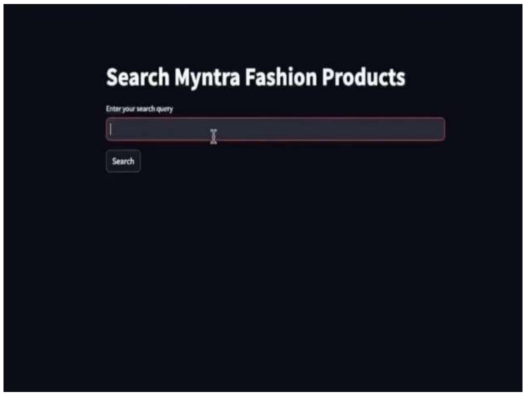
```python
df.head()
```

	ProductID	ProductName	ProductBrand	Gender	Price (INR)	NumImages	Description	PrimaryColor	DescriptionVector
0	10017413	DKNY Unisex Black & Grey Printed Medium Trolle...	DKNY	Unisex	11745	7	Black and grey printed medium trolley bag, sec...	Black	[0.027645884, -0.0026341688, -0.0035883961, 0...
1	10016283	EthnoVogue Women Beige & Grey Made to Measure ...	EthnoVogue	Women	5810	7	Beige & Grey made to measure kurta with churid...	Beige	[-0.02466071, -0.028755367, -0.020332502, 0.03...
2	10009781	SPYKAR Women Pink Alexa Super Skinny Fit High-...	SPYKAR	Women	899	7	Pink coloured wash 5-pocket high-rise cropped ...	Pink	[-0.046943244, 0.08182793, 0.048335165, -0.000...
3	10015921	Raymond Men Blue Self-Design Single-Breasted B...	Raymond	Men	5599	5	Blue self-design bandhgala suitBlue self-desig...	Blue	[-0.015098748, -0.010285424, 0.009487302, -0.0...
4	10017833	Parx Men Brown & Off-White Slim Fit Printed Ca...	Parx	Men	759	5	Brown and off-white printed casual shirt, has ...	White	[-0.0177466, 0.006209646, 0.021813976, 0.02679...

Figure 7. Indexed the Mapping about the dataset in the index

4.7.Streamlit UI Frontend Development

Developed a Streamlit based user interface for seamless interaction with the search engine and designed an intuitive frontend allowing users to query and visualize search results as shown in the Figure 8.

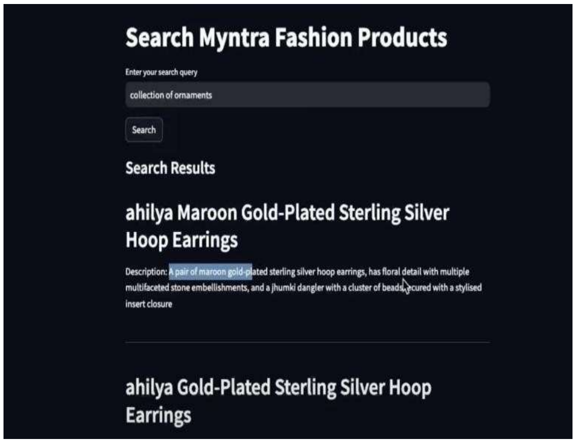

Search Myntra Fashion Products

Enter your search query

Figure 8. Frontend of the Search Engine

4.8.Quality Assurance and Testing:

Number tests conducted are 3 rigorous testing to validate the accuracy and functionality of the search engine as it gives results based on the context of query entered and addressed and resolved any issues identified during the testing phase as shown in the Figure 8.


Search Myntra Fashion Products

Enter your search query

collection of ornaments

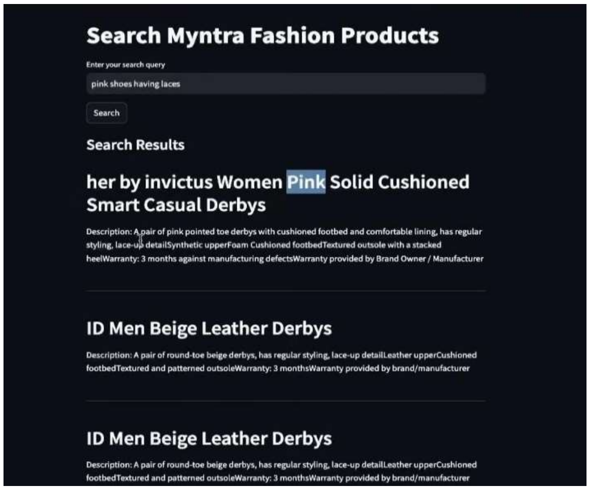
Search Results

ahilya Maroon Gold-Plated Sterling Silver Hoop Earrings

Description: A pair of maroon gold-plated sterling silver hoop earrings, has floral detail with multiple multifaceted stone embellishments, and a jhumki dangler with a cluster of beads, secured with a stylised insert closure

ahilya Gold-Plated Sterling Silver Hoop Earrings

Figure 9. Tested Context query search based on ornaments that give results about all the available data about ornaments in the dataset


Search Myntra Fashion Products

Enter your search query

pink shoes having laces

Search Results

her by invictus Women Pink Solid Cushioned Smart Casual Derbys

Description: A pair of pink pointed toe derbys with cushioned footbed and comfortable lining, has regular styling, lace-up detailSynthetic upperFoam Cushioned footbedTextured outsole with a stacked heelWarranty: 3 months against manufacturing defectsWarranty provided by Brand Owner / Manufacturer

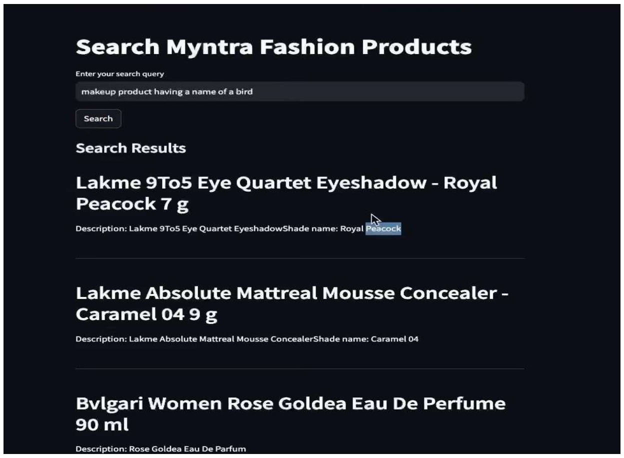
ID Men Beige Leather Derbys

Description: A pair of round-toe beige derbys, has regular styling, lace-up detailLeather upperCushioned footbedTextured and patterned outsoleWarranty: 3 monthsWarranty provided by brand/manufacturer

ID Men Beige Leather Derbys

Description: A pair of round-toe beige derbys, has regular styling, lace-up detailLeather upperCushioned footbedTextured and patterned outsoleWarranty: 3 monthsWarranty provided by brand/manufacturer

Figure 10. Tested Context query search based on colour (eg: Pink) that give results about all the available data about pink colour related things in the dataset


Search Myntra Fashion Products

Enter your search query

makeup product having a name of a bird

Search Results

Lakme 9To5 Eye Quartet Eyeshadow - Royal Peacock 7 g

Description: Lakme 9To5 Eye Quartet EyeshadowShade name: Royal Peacock

Lakme Absolute Mattreal Mousse Concealer - Caramel 04 9 g

Description: Lakme Absolute Mattreal Mousse ConcealerShade name: Caramel 04

Bvlgari Women Rose Goldea Eau De Perfume 90 ml

Description: Rose Goldea Eau De Parfum

Figure 11. Tested Context query search based on Peacock that give results about all the available data about peacock related things in the dataset

This implementation outlines a systematic approach to building a robust search engine architecture. Beginning with the setup of Elasticsearch and the creation of an index for CSV data, the process involves meticulous data processing with Pandas and integration of semantic vector embedding's using the SBERT model. The mapping of the index ensures compatibility with Elasticsearch, and the development of a Streamlit UI enhances user interaction. Rigorous testing and performance optimization contribute to a reliable and efficient search engine. This comprehensive methodology, coupled with ethical considerations and thorough documentation, lays the foundation for a scalable and user friendly information retrieval system.

5. Limitations

Even though the Elasticsearch based search engine has some good points, there are also things it can't do well. One big issue is that it's a bit tricky for people to learn how to set up and make it work perfectly. Users need to spend time understanding its details. Another limitation is that, even though the SBERT model helps understand the meaning of words, it might struggle with understanding some very specific details, making some answers not as accurate. Also, how well the system works depends on how good and big the dataset is. With smaller or less organized data, it might not work as well. These problems show that there's room to make things better, especially in making it easier to set up, understanding context better, and handling different types of data.

6. Future Works

Looking into the future, there are good ways to make the Elasticsearch search engine better. Firstly, we can make it easier for people to use by improving how it looks and making it simpler for new users to get started. This will help people who find it hard to set up Elasticsearch. We can also make the search engine smarter in understanding what users are looking for. This can be done by using more advanced techniques, like adding really smart computer programs that understand language better. This will make the search results more accurate when people type in their questions. Lastly, we can work on making the search engine work well with different types of information. Right now, it works well with some kinds of data, but we can make it work even better with all sorts of information. This will make the search engine more reliable and useful for everyone

7. Conclusion

In conclusion, the comprehensive development and assesment of the Elasticsearch based search engine have given understanding into its capabilities and identified potential avenues for refinement. The systematic approach, makes us understand about Elasticsearch, Pandas, and SBERT model embedding's, has yielded a search system characterized by remarkable speed, efficiency, and optimal resource utilization. The strategic integration of SBERT has notably elevated semantic understanding, contributing to enhanced search relevance. It is essential to acknowledge the challenges acquired with capturing nuanced contexts and the learning curve inherent in mastering Elasticsearch configuration. These findings underscore the need for ongoing efforts to address these challenges, ensuring continual improvement in the system's overall performance and user experience. The positive result of project besides being demonstrates the effectiveness of the chosen technologies but also sets the stage for further advancements and innovations in the realm of search engine development.

References

[1] Kamath S, Sowmya & Kanakaraj, Monisha. (2014). NLP based Intelligent News Search Engine using Information Extraction from E-Newspapers. 10.1109/ICCIC.2014.7238500.

[2] Mukhopadhyay, Debajyoti & Sharma, Manoj & Joshi, Gajanan & Pagare, Trupti & Palwe, Adarsha. (2013). Intelligent Agent Based Semantic Web in Cloud Computing Environment.

[3] Achsas, Sanae & Nfaoui, El Habib. (2022). Academic Aggregated Search Approach Based on BERT Language Model. 1-9. 10.1109/IRASET52964.2022.9737888.

[4] Patel, Manish. (2019). TinySearch -- Semantics based Search Engine using Bert Embed- dings.

[5] Rashid, Junaid & Nisar, Muhammad. (2016). A Study on Semantic Searching, Semantic Search Engines and Technologies Used for Semantic Search Engines. International Journal of Information Technology and Computer Science (IJITCS)International Journal of Information Technology and Computer Science(IJITCS). 10. 82-89. 10.5815/ijitcs.2016.10.10.

[6] Sadeeq, Mohammed & Zeebaree, Subhi. (2021). Semantic Search Engine Optimisation (SSEO) for Dynamic Websites: A Review. 5. 148-158. 10.5281/zenodo.4536804.

[7] Laddha, Shilpa & Jawandhiya, Pradip. (2017). Semantic Search Engine. Indian Journal of Science and Technology. 10. 1-6. 10.17485/ijst/2017/v10i23/115568.

[8] Mukhopadhyay, Debajyoti & Sharma, Manoj & Joshi, Gajanan & Pagare, Trupti & Palwe, Adarsha. (2013). Experience of Developing a Meta-Semantic Search Engine. 10.1109/CUBE.2013.38.

[9] Munarko, Yuda & Rampadarath, Anand & Nickerson, David. (2023). Building a search tool for compositely annotated entities using Transformer-based approach: Case study in Biosimulation Model Search Engine (BMSE). F1000Research. 12. 162. 10.12688/f1000research.128982.1.

[10] Sulistiyo, Edy & Wibawa, Setya chendra & Sujatmiko, Bambang & Nugroho, Dimas. (2021). The making of Android-based search engine applications with Elastic-search algorithm to improve programming competence. IOP Conference Series: Materials Science and Engineering. 1098. 042089. 10.1088/1757-899X/1098/4/042089.

[11] Malekar, Mrunal. (2021). Deep Learning-Based Question Answering Search Engine. In- ternational Journal of Scientific Research in Computer Science, Engineering and Information Technology. 25-32. 10.32628/CSEIT2172