© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716197

Smartdoc: AI-Powered Contextual Knowledge Assistant

SHILPA.K. S[1], PRATIBHA.H. G[2], SADHANA.S. R[3], SAHANA.D.M[4], SONIYA KOMAL[5]
[1,2,3,4,5]*Department of Computer Science Engineering, Rajiv Gandhi Institute of Technology, RGC Campus, Cholanagar, RT Nagar Post, Hebbal, Bangalore*

Abstract- ***The increasing amount of unstructured data has posed problems with respect to their proper processing and retrieving information from them. Various attempts have been made by researchers to find answers through AI and NLP approaches such as document summarization, speech-to-text conversion, and semantic search techniques. Yet, the existing approaches are limited to single-source approaches, meaning that none of them is capable of dealing with multiple sources. In this paper, we will use Smart Doc: AI-Powered Contextual Knowledge Assistant, which can be considered as an integrated approach for the processing of multiple sources of data, such as documents, video transcripts of YouTube videos, and comments. We leverage the potential of semantic embedding and the RAG technique to process multiple sources of data. Smart Doc demonstrates high accuracy and effectiveness compared to traditional methods based on keyword matching.***

Index Terms- ***Retrieval augmented generation (RAG), Natural Language Processing, YouTube transcript analysis, Semantic search***

I. INTRODUCTION

In today's world, knowledge is spread across a variety of media, including documents, research papers, and multimedia channels like YouTube. The issue with the above data sources is that there is no structure, and the knowledge available is disorganized and disjointed, which makes it difficult for users to retrieve insights easily. Most search tools work based on keyword searches, which have limited capacity when it comes to understanding the actual context behind the queries made by users. Because of the above limitations, users often find themselves forced to go through entire documents or video transcripts to gain access to relevant information. In addition to this, most of the platforms lack a mechanism that can bring together multiple data sources, including transcripts of video platforms like YouTube, comments of people, and even written documents. Lack of integration results in inefficiency when it comes to knowledge extraction to derive actionable insights. To overcome the above problems, we propose Smart Doc – an AI-powered contextual knowledge assistant that leverages the latest advances in artificial intelligence.

II. NEED OF THE STUDY

In today's digital age, people are constantly bombarded by huge volumes of data from various sources like documents, online video tutorials, user comments, and discussions. Even though a platform like YouTube generates a massive amount of content every minute, it becomes really challenging for users to find the exact thing. Hence, people end up spending their time reading long documents, watching whole videos, or looking through comments to find out just one concept. The traditional search systems work on the concept of keyword matching, which is unable to understand the essence of a particular search term. It makes the gap between users' intent and results more apparent. There is an urgent necessity for an advanced approach that can analyze the context of users' search, identify its main idea, and give results accordingly. Thanks to the progress of technologies in AI and NLP, it has become possible to develop more intelligent systems for processing human language. At the same time, most of the available solutions focus on working with one type of data at once, regardless of its source, whether it is documents or videos. In this case, there arises a requirement for a new system. An equally important aspect involves the problem of non-integration of different sources of data. Current systems consider texts, transcripts, and comments in isolation, which makes it impossible for users to have a holistic view of information. Additionally, language models may provide answers based on general knowledge without accessing the particular data of the user, thereby creating hallucinations.

The need to develop smart solutions that minimize human effort, enhance the search process, and offer context-sensitive insights is becoming increasingly common. In the absence of these technologies, it becomes necessary to create an integrated platform to extract knowledge from diverse sources of data.
Hence, this research is crucial as far as the development of an advanced AI system capable of addressing the above limitations is concerned.

2.1 Population and Sample
In the context of this study, population can be defined as the total number of information resources and users that will be used in order to examine the Smart Doc system. In general, the population will consist of unstructured data sources, which will be used to generate useful information. These unstructured data sources will be in the form of documents, transcripts of YouTube videos, and user comments. Such data will be taken into consideration because of the fact that such unstructured data sources are generally available in reality. In other words, such data is commonly used by students, researchers, and professionals.

A sample will be derived from such a population to perform the examination of the performance and efficiency of the Smart Doc system. In addition to data sampling, a sample group of users is chosen to interact with the Smart Doc system. These users will generate queries and assess the performance of the system according to various criteria, including accuracy, relevance, and speed. The data collected from these users will be useful in measuring user satisfaction and system usability.

The sampling strategy employed for this research may be regarded as a blend of purposive sampling and random sampling methods. It will help in selecting an appropriate sample that is suitable and representative of the Smart Doc system.

In general, the described population and sample are instrumental in assessing the performance of the Smart Doc system.

2.2 Data and Sources of Data
The main source of data consists of documents in the form of PDF files and scholarly articles, where detailed and organized information is present for different subjects. Besides, video data is obtained from websites such as YouTube, and transcriptions are extracted from videos using APIs that facilitate the conversion of spoken language into text. Document data, Transcript data, and Comment data are preprocessed before their utilization in the Smart Doc system. The processes involved include cleaning, normalization, and conversion of the data into embeddings with the help of tools such as sentence-BERT. After preprocessing, the data is loaded in the vector database to allow easy semantic search.

2.3 Theoretical Framework
Smart Doc technology is founded on a theoretical framework that brings together ideas related to AI, natural language processing, and semantic information retrieval. Summarization is part of the theoretical framework of the Smart Doc technology, where huge amounts of information are reduced to short outputs. Comment analysis is another aspect incorporated in the theoretical framework of the Smart Doc technology.

III. RESEARCH METHODOLOGY

The research methodology used in this research is oriented towards designing a system that can be effective in analyzing unstructured multi-source data to draw out significant information from it. Artificial intelligence and natural language processing methods have been used in the research methodology. The research design is based on the development and experimental testing of the model under study. The system is intended for solving practical tasks associated with the processing of different types of data through the use of innovative AI approaches.
.
3.1 Population and Sample
The population in this case study includes unstructured data sources and the interaction between users and digital knowledge management systems. The population of data includes documents, transcripts from YouTube videos, and comments generated by users. This represents actual data that would be found in academic, research, and professional settings. The population of users includes students, researchers, and professionals who

regularly interact with large amounts of text and multimedia content and need efficient information search tools. A sample is taken from the population to examine the effectiveness of the Smart Doc system. The sample consists of selected documents, transcripts from YouTube videos, and user comments on various subjects. The sample is selected in such a way that it reflects diversity and can be used to test the system under realistic circumstances. Moreover, the interaction of users with the system in the form of queries and responses is also included as part of the sample. The sampling frame consists of unstructured data sources like documents, transcripts of videos on YouTube, and comments from people. Other sampling units include users, including students, researchers, and experts who access knowledge systems.

A random sampling technique is adopted for choosing a few documents, transcripts, and comments from different fields. User input and responses are analyzed to measure system efficiency.

3.2 Data and Sources of Data

The Smart Doc model relies on datasets obtained from diverse sources to allow knowledge discovery.
The sources of datasets can be summarized as follows:

- Documents: Research articles, PDFs, and text files
- Video transcripts: Generated through the use of the YouTube API
- User comments: User-generated content that offers additional context

It should be noted that these types of datasets provide different types of unstructured data and play an important role in creating a universal knowledge base. The next stage consists of preprocessing, which encompasses activities related to text cleaning, normalization, tokenization, and elimination of unnecessary elements. In the following step, the preprocessed dataset will be converted to semantic vectors using algorithms, for instance, sentence-BERT. Finally, the processed dataset will be stored in a vector database.

3.3 Theoretical Framework

The theoretical framework for the Smart Doc application is founded on semantic information retrieval and context-based knowledge acquisition. The application employs embedding-based modeling, whereby text data is represented in a high-dimensional vector space through techniques like sentence BERT. This technique ensures that the system can interpret the context of the text rather than matching the keywords only. The theoretical framework integrates a vector database capable of storing embeddings and performing searches using similarity. Whenever a query is entered by a user, the system represents it in a vector format and compares it to existing data to provide the relevant output. The system is founded on semantic information retrieval and embedding-based representation. Textual data is encoded into vector representations by means of Sentence-BERT models, among others.

An essential element is the Retrieval-Augmented Generation method, which integrates both retrieval and generation to ensure high-quality, context-relevant outputs. The theoretical foundation of the proposed Smart Doc intelligent system lies in the combination of theories related to Artificial Intelligence and Natural Language Processing. These two fields form the basis for constructing systems that will be able to process the language, its meanings, and generate content accordingly. It is the basis of creating models that transform unstructured information into structured knowledge.

One of the essential principles of such an approach is semantic representation of information, where textual data is translated into numbers or vectors that will represent the semantic content of the input. Contrary to simple keyword searches, the proposed system will allow searching for similar sentences regardless of their wording or even structure. This will be achieved by the usage of Sentence-BERT, which will be responsible for the vectorization of texts.

All documents, transcripts, and user comments will be saved in the vector space database, which means that once the user enters a query, its corresponding vector will be generated and compared with other vectors through the calculation of similarity measures.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716197

IV. STATISTICAL TOOLS AND ECONOMETRIC MODELS

Statistical and econometric methods are used in this research to interpret the results as well as assess the accuracy of the model. Some of the methodologies employed in the research include the use of descriptive statistics, a regression model, and a comparison of models.

4.1 Descriptive Statistics

Descriptive statistics help greatly in defining and describing the basic nature of the data used for this study. The performance of the Smart Doc system is described using measures like accuracy, response time, relevance score, and user satisfaction. The use of descriptive statistics in Smart Doc is important because of its usefulness in providing insight into the performance of the system.

It is important to point out that descriptive statistics serve the important function of simplifying data into meaningful value measures. Descriptive statistics include mean, median, and mode. It can be noted that the mean will indicate the accuracy of the system in performing functions. In addition, the median will reduce the impact of outliers on the result and indicate how consistent the results of the Smart Doc system are. Apart from central tendency, other measures, for instance, standard deviation and variance, are used to measure the spread of the data in terms of system performance. A low standard deviation shows that the system consistently performs, while a high value indicates inconsistent system performance. This is especially significant when it comes to assessing the reliability of the system with respect to different sources, such as transcripts and documents uploaded via YouTube.

Descriptive statistics provide information on other measures, including minimum and maximum. The minimum provides information on the fastest response of the system, while the maximum gives the system response under complicated scenarios. Therefore, with these pieces of information, we get an understanding of both the best-case and worst-case scenarios of the system.

Also, through descriptive analysis, we can obtain useful information regarding the trends, patterns, and even outliers of the data. This can be critical in helping us understand unusual cases and any errors in the data processing that might impact system performance.

4.2 Fama-MacBeth Two-pass Regression

The Fama–MacBeth two-pass regression is a statistical approach that is popularly applied in estimating the connection between dependent and independent variables, while at the same time taking into account problems like time variability and cross-sectional correlation. Although initially designed for financial studies, the approach can be modified for analytical purposes by assessing the effect of various factors on system performance in numerous instances.

The process comprises two separate steps. The first step involves conducting time series regressions for each individual observation unit. For the case of the Smart Doc system, the first step may be understood as testing the variations in performance indicators, including accuracy, response time, and relevancy score in varied data sets or periods. The goal of the first step is to establish sensitivity coefficients (or betas). In the second run, cross-sectional regression is applied based on the coefficients gained from the first run. At this stage, the effect of different factors on system performance is estimated using the comparative analysis of several samples. Thus, for the case of Smart Doc, one may consider the contribution of semantic search and embedding models, as well as the role of Retrieval-Augmented Generation.

The main strength of the proposed technique lies in its capacity to eliminate the estimation bias, which may appear due to the application of single-stage regressions to analyze the influence of factors on system performance. With the help of the suggested approach, one takes into account the heterogeneity of the results due to variations in time and data used (document transcripts from YouTube, for instance).

4.2.1 Model for CAPM

The Capital Asset Pricing Model (CAPM) is one of the most popular models in finance that describes the

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716197

correlation between the expected returns and systematic risk. The model suggests that the expected returns of assets are dependent on their sensitivity to the overall riskiness of the stock market, which is described using beta. The CAPM is rooted in the assumption that the returns of investments should only be compensated for risks that cannot be diversified away, but not the other way around.

The Capital Asset Pricing Model(CAPM) explains the relationship between expected return and market risk

E(Ri) = (Rf)+beta(Rm-Rf)
E(Ri) = Expected return of asset
Rf = Risk-free rate
Rm = Market return
Beta = Sensitivity to market risk
CAPM assumes that market risk is the only factor influencing returns.

As far as the Smart Doc case study is concerned, the CAPM equation may be modified in order to assess the influence of one particular influencing element on the performance indicators, such as accuracy or performance. In this case, “the market rate of return” will stand for the general system performance indicator, and “beta” will measure the sensitivity of the system to certain influencing elements. While this equation is not able to assess multi-factor influencing on the system, it still gives a clear idea about how certain elements influence the system’s performance.

4.2.2 Model for APT
Arbitrage Pricing Theory (APT) is a multivariate asset pricing theory that describes the connection between the expected return on assets and several risk factors. Contrary to CAPM, APT takes into account several risk factors because in APT theory, it is presumed that the returns depend not only on a single risk factor but rather on several risk factors that include but are not limited to inflation, interest rates, economic development, and growth, etc. In addition, APT is underlined by arbitrage – the idea that there can be no riskless profit opportunities in a market.

$E(Ri)=Rf+\beta1F1+\beta2F2+\beta3F3+……..+\beta nFn$
Where:
E(Ri)=Expected return
Rf=Risk-free rate
F1,F2,…….,Fn=Different risk factors
B=Sensitivity of the asset to each factor

In relation to the Smart Doc research study, the APT framework is an appropriate theoretical framework that could be used to measure how multiple variables affect system performance. The variables in question could include semantic search speed, embedding quality, multiple sources integration, and Retrieval-Augmented Generation performance. Each variable is important in the determination of overall system performance metrics, including speed and relevance of responses and information provided by the system. Additionally, through APT, researchers can measure which of the variables contribute more significantly to the overall effectiveness of the system. For instance, it could be possible to measure which aspect affects the accuracy of the information provided, for instance, whether it is more effective to improve embedding models than data pre-processing. Overall, APT framework is an appropriate approach to analyze the system.

4.3 Comparison of the Models
The performance of CAPM and Apt models is compared based on explanatory power, flexibility, and accuracy.

- CAPM is simpler and easier to implement
- APT provides better explanation using multiple factors
- APT generally offers improved predictive performance

Model	Type	Factors considered	Accuracy	Flexibility	Suitability for Smart Doc
CAPM	Single-factor	Market risk only	Moderate	Low	Not suitable
APT	Multi-factor	Multiple risk factors	High	High	Limited use
Keyword Search	Traditional	Exact word matching	Low	Low	Poor

TF-IDF	Statistical	Team frequency	Medium	Medium	Limited
Embedding Models	Semantic	Context meaning	High	High	Good
Generative AI models	AI-based	Language patterns	Very High	High	Good
RAG-based Model	Hybrid	Retrieval+ Generation	Very High	Very High	Best
Smart DOC(proposed)	Hybrid AI	Multisource +Context	Flexibility	Very High	Optimal

4.3.1 Davidson and MacKinnon Equation

Davidson–MacKinnon Test refers to a type of econometric technique that helps in choosing between two competitive models in such a way that it tells about which model best fits the given set of observations. This technique differs from the other existing techniques of comparing two models in the sense that there is no need for nested models for testing.

The general form of the Davidson and Mackinnon equation is

$$Y=\alpha+\beta X+\gamma \hat{Z}2+\epsilon$$

Where:
Y=Dependent variable
X=Independent Variable
$\hat{Z}2$=Predicted values from the competing model
γ=Coefficient used to test model validity
ϵ=Error term

If the value of γ turns out to be statistically significant, then it shows that the second model explains the relationship better and may be regarded as better in comparison. On the other hand, if it does not show statistical significance, the first model will suffice.

From the viewpoint of the Smart Doc project, the application of this test enables comparing different techniques, such as classical keyword searching and an AI-based one. In particular, outputs from one type of model may be predicted and introduced to the second one (Smart Doc), in order to prove the fact that the latter works much better. Thus, it becomes possible to state that Smart Doc, being a complex of semantic search and Retrieval-Augmented Generation, demonstrates better results than traditional models.

As a result, the Davidson-Mackinnon test becomes a powerful instrument to conduct model comparison and to find the best of them in accordance with statistics. The effectiveness of the test enhances the validity of the study.

4.3.2 posterior Odds Ratio

Definition

The Posterior Odds Ratio is a ratio employed in Bayesian inference to compare two models that compete with each other through probabilities that result from the observation of some data. The Posterior Odds Ratio serves as a quantitative method to identify which of the models is more probable to be true according to the evidence. In comparison with the usual method of hypothesis testing, the Posterior Odds Ratio can be considered even more effective because it allows comparing models.

It is defined as the ratio of posterior probabilities of two models: Value>1→First model preferred, Value<1→Second model preferred
This method is commonly used in Bayesian model selection

Formula:- $P(M1/D) / P(M2/D)$

Where:
P(M1/D)=Probability of Module 1 given data
P(M2/D)= Probability of Module 2 given data

These values are calculated based on the use of Bayes' theorem, according to which the posterior beliefs about models should be updated according to some new data. The Posterior Odds Ratio is larger than 1 when Model 1 is more preferred because of the available data. If this value is less than 1, then Model 2 is more probable. A ratio of 1 means that both models are equally plausible.

Applying the concept of Posterior Odds Ratio to the Smart Doc system, we can evaluate different approaches to the search for information and its

generation. So, Model 1 is the system including semantic search and Retrieval-Augmented Generation in it, while Model 2 can be the keyword-based system that does not rely on semantic search. Using some performance parameters such as accuracy, relevance, and response speed, we can assess which approach performs better.

V. RESULTS AND DISCUSSION

Analysis of the results obtained through the application of the Smart Doc system is done using descriptive statistics and performance measurement tools. The purpose of this analysis is to determine the efficiency of the system in terms of accuracy, speed, relevance, and customer satisfaction. The results have been represented in tables for easy understanding.

5.1 Results of Descriptive Statistics of study variables

Table 5.1: Descriptive statistics

Variable	Minimum	Maximum	Mean	Median	Standard Deviation
Accuracy(%)	85.0	97.0	92.5	93.0	3.2
Response Time(S)	1.0	3.0	1.8	1.7	0.5
Relevance Score	0.80	0.95	0.89	0.90	0.04
User Satisfaction	3.8	5.0	4.5	4.6	0.3

The above-presented table describes the main characteristics of the performance of several key performance indicators that allow us to evaluate the Smart Doc system. All these metrics help us assess the effectiveness of using the system for processing queries and providing users with the information needed.

In particular, the high value of the mean of the accuracy metric (92.5%) proves that Smart Doc performs really well and returns only the relevant data. In addition, the small difference between the mean (92.5%) and the median (93.0%) indicates that the distribution of this indicator is rather balanced, while a relatively low standard deviation (3.2) means no fluctuations in the results received.

Another interesting metric characterizing performance is the mean response time. According to the presented figures, it takes 1.8 sec. The small difference between the minimal (1.0 s) and maximal (3.0 s) response times means that the Smart Doc system works stably even during processing complex queries.

Finally, we should pay attention to the relevance score (mean value equal to 0.89). This high figure means that the obtained results can be considered appropriate. As a result, the use of semantic search and Retrieval-Augmented Generation has been proven to work properly.

VI. ACKNOWLEDGEMENT

The authors would like to convey their heartfelt gratitude to all those who made this research work possible by contributing to its successful completion. First of all, we would like to convey our gratitude to our project guide for their continuous guidance, helpful suggestions, and support in the process of developing this project.

We would also like to thank our institution for providing us with the necessary infrastructure as well as academic support in carrying out this research project.

We are thankful to various websites and online platforms like YouTube, which helped us in obtaining transcriptions from various video recordings for implementing our software project.

At last, we would like to thank our peers as well as other individuals who assisted us in carrying out this research project.

REFERENCES

[1] A. K. Maiya, G. S., A. R. V., and N. S. C., "YouTube Transcript Summarization Using Abstractive and Extractive Approaches," Proceedings of the 2024 2nd International Conference on Advances in Information Technology (ICAIT), 2024.

[2] "Paper presented at ICISS 2025", Proc. 2025 Int. conf. on intelligent systems and smart systems, 2025.doi:10.1109/ICISS63372.2025.11076484

[3] S.A.A Rizvi, S.R. Javed, and M. A. Hashmani, "Document AI: Benchmarks, models and applications," arXiv preprint arXiv:2111.08609, 2021

[4] S. Tay., M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," arXiv preprint arXiv:2009.06732,2020