EasyChair Preprint

№ 11160

Survey : Study of the Improvement of IR and SEO for a Better E-Reputation

Djoudi Kaouthar, Alimazighi Zaia and Dellal-Hedjazi Badiâa

October 24, 2023

Survey: Study of the improvement of IR and SEO for a better e-reputation

***Abstract*— The exponential increase in the amount of information, its variety and its velocity requires addressing them with new, more powerful tools for accessing relevant information in a short time. As well as the birth of the concept of e-reputation on the Web this one requires the improvement of the content of sites for better referencing by search engines. Our article is the state of the art on the research of classical techniques of artificial intelligence for improving semantic search. Such as latent semantic analysis (LSA), probabilistic latent semantic analysis (pLSA) or their improvement the Latent Dirichlet Allocation (LDA). And the new techniques such as deep learning or more recently Transformers (BERT, GPT, … etc.) for improving information retrieval (IR) and search engine optimization (SEO) by integrating semantic aspect in order to improve the natural referencing of a company or for a better e-reputation of the company in the Web. So we study the techniques used in the most recent works on several aspects, such as the accuracy and relevance of the information, the performance and quality of the result and the speed of obtaining the information. And we are doing a comparative study on their benefits and limitations. In order to show which technique is the most effective for guidance in our future work**.

***Keywords*—** *IR, SEO, LSA, pLSA, LDA, Deep learning, Transformers, e-reputation*

I. INTRODUCTION

In recent years the web has seen an exponential increase in the amount of information. Moreover, it saw a wide variety of heterogeneous data such as structured, semi-structured and unstructured data as well as the velocity or speed of data collection and loading. These different factors require addressing them with new, more powerful tools for access to relevant information with semantics taken into account and in a reduced time. On the other hand, the birth of the concept of e-reputation or the reputation of the company on the website requires new tools, methods or techniques for improving SEO (Search Engine Optimization).

There is a lack of fast and efficient tools to search for information on Big Data to answer queries with relevant results. Also, there is a lack of Big Data content optimization tools for improving the natural referencing of sites. Particularly on their semantic aspects, requires the use of artificial intelligence techniques, whether classic (LSA, pLSA or LDA) or new (Deep learning or recently Transformers) or combine the two to create new methods, tools or techniques to improve IR and SEO for a better e-reputation.

Our objectives are first to study the different AI techniques to used them in our future work to improve retrieval information and SEO in relevance and velocity taking into account semantics and treating the Big Data.

The optimization of web content by content classification, the removal of unnecessary publication and the use of tools for the best management of the website are SEO factors that improve the company's e-reputation. As well as the methods and techniques which make it possible to recover the most relevant results to the user's request in a precise time allowing to improve the natural referencing of the site or to improve the e-reputation of the company. This is why our article is the state of the art on improving information retrieval (IR) especially in the semantic aspect and improving content so that it can be easily repaired by search engines (SEO).

We perform these improvements using classic techniques of artificial intelligence such as latent semantic analysis (LSA), probabilistic latent semantic analysis (pLSA) or their improvement Dirichlet allocation (LDA) and new techniques like deep learning or more recently Transformers (BERT, GPT…etc.). In the first part we will define and explain them. In the second part we will present research works that have dealt with these areas and study how they have used it in order to improve IR and SEO in relevance and speed. At the end, you will find our research orientation, some ideas for future work and then a conclusion.

II. BASIC CONCEPTS

A. *Information Retrieval (IR)*

Information retrieval (IR) consists of finding the relevant result content according to the user's information needs through his query formulated from large collections (Big Data) [1]. In recent years, the development of natural language processing (NLP) and neural networks has considerably enriched information retrieval methods [2].

B. *Basic IR process system (IRS)*

The information retrieval system (IRS) is a system that allows to process information and interact with the user, it can be schematized in Fig.1 [9] [10].

When the user's query arrive, the IRS retrieves from the inverted index table the most relevant documents for the query [10]. The IRS index is called an inverted index because normally documents are stored as lists of words, but inverted indexes reverse this by storing for each word the list of documents in which the word appears [34]. Then the IRS sends them as candidate documents using TF IDF (Term Frequency-Inverse Document Frequency) or BM25 (Best Matching 25) scoring. Finally, these documents go through

the stage of classification from more to less relevant in order to send them back to the user [10].

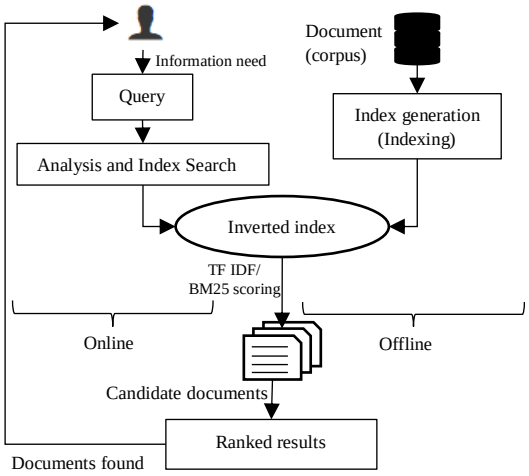

Fig.1.Basic Information Retrieval System Process

C. *Search Engine Optimisation (SEO)*

Search Engine Optimization is a technology that shows how search engines perform searches and how to determine the ranking of search results for specific keywords by analyzing search engine ranking rules. These compare the contents of web pages and then provide users with the most complete, direct content through the browser without affecting the user experience. Therefore, SEO is the set of methods used to improve the natural referencing of websites. Its main task is to understand how to index, determine search keywords and other related technologies in order to optimize web content, index it and rank it well in search engines [3].

D. *Latent Semantic Analysis (LSA)*

Latent semantic analysis (LSA) is one of the fundamental techniques of NLP (natural language processing) and IR [24] [25]. It makes it possible to extract the meaning of words by statistical calculation applied to a vast corpus of texts [6] generally unstructured data [6] [23]. The steps of an LSA process are:

1. Construction of the document-term matrix (A) [26]

   Build A where each row represents a document and each column represents a term and each element of A represents the frequency of the term in each document, where each word has TF and IDF scores and the product of these (TFIDF) is the weight of the word.

2. Reduction of the dimensions of A [4]

   Matrix A has a lot of redundancies, noises and dimensions. In order to keep the most significant words and to capture the latent subject that describes the relationship between the words and the documents, A reduces its dimensions using SVD (singular value decomposition). This linear algebra technique factors the matrix A into the product of three matrices: A = U*S*V. Where U is the document-subject matrix, S is a diagonal matrix of the singular values of A and V is the term-subject matrix.

3. Studying cosine similarity [4]

   Using these three matrices, the cosine similarity calculation is applied to evaluate: the similarity of different documents, different words and the similarity between queries and documents, which becomes useful in IR to retrieve the most relevant passages for the request.

LSA is fast and efficient to use, but it has a few main limitations [4]:

- Lack of interpretable incorporations (subjects are unknown and components can be arbitrarily positive/negative).
- Need a very large set of documents and vocabulary to obtain precise and relevant results.
- Less effective representation.

E. *probabilistic Latent Semantic Analysis (pLSA)*[4]

pLSA uses a probabilistic method to find the latent semantics instead of the SVD (used for LSA). In principle it creates a model p(d,w) (d document and w word) indicates the probability of seeing a document d (p(d)), then, depending on the distribution of subjects z of this document p(z |d), the probability of finding certain word w in this document p(w|z). As shown in equation (1) and simpler in equation (2). d is determined directly, p(z|d) and p(w|z) are obtained from the expectation maximization (EM) algorithm. It is a method to find the most likely parameter estimates for a model that depends on latent variables. The probability model of pLSA is modeled in Figure 2.

$$p(d,w) = p(d)\sum_z \mathrm{p(z,d)}\, p(w,z) \quad (1)$$

$$p(d,w) = \sum_z \mathrm{p(z)p(d,z)}\, p(w,z) \quad (2)$$

There is the parallel between LSA and pLSA, where p(z) corresponds to the matrix A, p(d|z) to the matrix U and p(w|z) to V. However, pLSA adds a probabilistic treatment compared to LSA. It has a more flexible model while still having some problems like:

- The absence of parameters to model p(d) causes the inability to assign probabilities to new documents.
- The number of pLSA parameters increases with the number of documents linearly, so it is subject to overfitting.

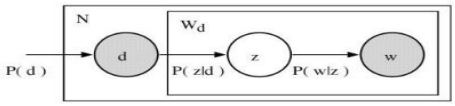

Fig.2.Probabilistic Latent Semantic Analysis model

F. *Latent Dirichlet Allocation (LDA)*

LDA is a Bayesian version of pLSA which is based on distributions over distributions. The pLSA model makes it possible to extract subjects from a corpus of text without specifying which subject is dominant, however the LDA model is more specific where each subject is parameterized by a weight, so it makes it possible to specify the dominant subject in a corpus of text. Thus, the LDA model can be easily generalized to new documents [4]. It assumes that each text corpus is a mixture of latent topics and each topic has a probability distribution over all vocabulary words [28, 29]. The principle of the LDA process is from a first Dirichlet distribution Dir(α) a random sample of topic distribution is

drawn from a particular document (θ), from θ a particular topic Z is selected based on the distribution. Then from

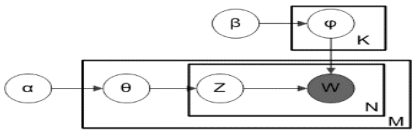

Fig.3.Latent Dirichlet Allocation model

another Dirichlet distribution Dir(β) a random sample is selected representing the word distribution of subject Z(φ). From φ, the word W is chosen, it is represented in figure 3 [4].

G. Deep learning

Deep learning is a subfield of machine learning [5]. Which is a form of artificial intelligence allowing a system to learn without being explicitly programmed [6]. The process of deep learning start by the input data that assigned for training the model to implement a series of transformations. Then the loss function controls the output by taking the network's predictions and the actual expected target and calculates a distance score. Naturally, its output data is far from what it ideally should be, and hence the loss score is very high. This is the training loop that repeats itself a sufficient number of times producing weight values that minimize the loss function. This adjustment is the work of the optimizer which implements what is called the back propagation algorithm. The fundamental trick of deep learning is to use this score as a feedback signal to slightly adjust the value of the weights in a direction that will reduce the loss score. Now, the trained model is ready to use it on data not seen during training [5].

H. Transformers

Transformers were proposed by [8] in 2017. They are a neural network that learns the context. Their models are the most powerful to date thanks to their application of a set of mathematical techniques, called attention or self-attention [07]. They are based on the encoder-decoder architecture that appeared by Cho, K., Van Merriënboer, B., Gulcehre and all in 2014 [30]. Its principle is the joint use of two modules, the encoder and the decoder. The encoder is a network of recurrent neurons that transform a sequence into a vector representation. And the decoder makes it possible to transform the output of the encoder into a final output [31].

Transformers are very efficient but have some disadvantages such as: Expensive to learn, they exploit a very large number of parameters and therefore take a long time [32] [33].

According to our research we list the main Transformers that handle natural language processing (NLP) in Table 1

In the next section of the article, we begin by quoting the works that improve IR and then works that improve SEO using AI techniques with explaining their methods.

TABLE I. TRANSFORMERS PROCESSING NLP

Transformer	**Description**	**Type**	**Among their uses**
BERT [20]	Bidirectional Encoder Representations from Transformer	Encoder, bidirectional	Classification
GPT [21]	Generative Pre-trained Transformer	Decoder, unidirectional	Text generation
T5 [22]	Text To Text Transfer Transformer	Encoder-Decoder	Translation, Question answering

III. RELATED WORKS

A. Improving IR using AI techniques

Kaveti Naveenkumar and shrutendra harsola in 2020 improved the basic IR system process, they touch the semantic aspect using Transformers (BERT). At the creation of the user's request, they added two steps: the incorporation of request and the incorporation of documents (They converted the request and the documents into digital vectors) using BERT. In order to study the similarity between the two incorporations. But they found that calculating the cosine similarity between all the embeddings is too expensive so they used ready-made libraries like FAISS to find the nearest neighbor to the document index and retrieve them [10].

Samarth Rawal and Chitta Baral in 2020 introduced the concept of a multi-perspective IRS using BERT to improve the IR system. The first step of their system is to take the user's query and retrieve the k best documents for this query on the large dataset (k is an integer defined by the user according to his objectives). The second part of classification takes the documents, decompose them into sentences, and applies the following three BERT models [13]:

- BERT Sentence relevance: Training the BERT Large model for a binary classification task (query-sentence) into a "relevant" or "irrelevant" pair in order to specify the degree of relevance.
- Semantic Textual Similarity (STS) : is a reference that evaluates how similar two sentences are (it is useful for synonyms)
- Semantic Information Availability (SIA): It answers the following question: How much information does a sentence have that is needed to answer a given query?

Finally, the first n most relevant sentences identify themselves, (n is an integer defined by user). From their research, they found that the combination of traditional keyword-based approaches and modern like transformers based approaches have proven effective in recent work. For ranking, BERT models are very computationally expensive compared to ranking algorithms like BM25, so it is better to use BERT for reranking to refine the model [13].

ORMEÑO, Pablo and all in 2021 used LDA for the IR Ad hoc; it is a document classification for a query by BOW (Bag Of Words). Their goal is to improve relevance by solving polysemy and monosemy problems caused by classical IR methods which are based on text, query and

document vectors. Their model uses the Bagging strategy which splits the text corpus into m sub-corpus partitions at random, fitting an LDA model in each partition (LDA Ens). Then, given a BOW query, they produce a ranking list from each model. Finally, they constructed a consolidated list by merging the rank lists obtained from each LDA model using the CombMNZ list merging method. For classification, boosting (LDABoost) is better than BOW in accuracy [12].

YANG, Nakyeong, JO, Jeongje, JEON, Myeongjun and all in 2022 concatenated both LDA and BERT techniques to create a recommender system. By taking the result of BERT (the sentence representation) with the result of the LDA process (the topic representation) to go through a sentence embedding layer then an output layer that predicts the result. Based on their evaluation, they found that using BERT + LDA gives better results in measurement of Accuracy, Precision, Recall, Score and Auroc. Also, the loss function during training and model testing gives inferior results [11].

KIM, Su Young, PARK, Hyeonjin, SHIN, Kyuyong and all in 2022, modeled a three-step GPT-3-based product recovery system [14]:

- Retrieval model: Retrieving the query and outputting the top k most relevant categories using GPT-3.
- Category to product mapping table: Categorize the results of the first stage using a mapping table in order to prepare the candidates.
- Ranking model: Classification of results using the BERT ranking model with the MLP multilayer perceptron, it allows learning the latent semantics of the query and words and calculating the score of the similarity between words and queries.

They found that the knowledge base stored in GPT-3 helps give superior performance over BERT with fewer parameters. Either for product recovery or for the general case of IR. Although the size of the model has an influence on its performance (the higher the size, the higher the performance) [14].

BARONETZKY, Pia, NIGMATOV, Noir, STOICAN, Theodor and all in 2021 developed a semantic search engine to index a very large number of documents. Their goal is to put similar texts to the query in context and semantics to return the most relevant documents for the query using BERT and LSA. Their model included Sentence-BERT, which is specifically pre-trained for semantic textual similarity tasks, BERT Extractive Summarizer, Latent Semantic Analysis (LSA), and LexRank. Comparing these last three for text summarization, they found that LexRank is fastest then LSA then BERT Extractive Summarizer. Because BERT and RoBERTa perform the similarity calculation by comparing sentences in pairs to obtain a similarity score. This leads to a significant computational overhead. So BERT is not suitable for calculating semantic similarity for a very large data set because it is very heavy and unfeasible in performance time. So they chose to do the text summarization with DistilBERT which is 60% faster than BERT and preserves 95% of the accuracy of BERT. Then they used [15] :

- GPT-2 for embedding, it's very slow because of those huge settings but it's the best for text generation.
- Pegasus model for abstract text summarization, its structure consists of an encoder and a decoder
- T5 because it achieves cutting-edge results in semantic similarity computation tasks when comparing sentences in pairs although it is slow.

Then, they used the average vector pooling of the previous three models to produce a single vector. Finally, they used FAISS to accelerate the prototyping and testing phases. According to the experiments of Reimers and al, the direct use of BERT leads to very low performance. And based on the comparisons made they chose BERT Sentence (SBERT) for scalable semantic search. SBERT is a modification of BERT/RoBERT that is able to derive semantically meaningful sentence embeddings [15].

The main limitation of classical IR techniques based on keyword indexing is to capture the semantics of contexts with a strong similarity, written with different words (synonyms). However, the cosine similarity metric is used to compare documents in the high-dimensional integration space. By using keyword indexing, it is possible to include or exclude specific terms in the query. So they concluded that the weighted linear combination of topic modeling and Sentence-BERT integrations allows to see different levels of semantic similarity and very general search results [15].

NOGUEIRA, Rodrigo and CHO in 2019, described a simple implementation of BERT, for the reordering of text passages based on a query. They have improved the redeployment phase of an IRS. They took MS MARCO which is a system that contains a million queries from real users and their respective relevant passages annotated by humans, and they repurposed BERT to reclassify their passages for their queries to replicate publicly available experiences [16].

Table 2 examines each of the works that improve IR by mentioning the techniques used in the work, these benefits and their limits.

B. Improving SEO using AI techniques

Anastasiu, C., Behnke, H. and all in 2021 addressed one of the main factors of improving e-reputation, which is the title of search results. Their model called BERTSUMABS creates as follows: First, they filter the article and title sizes. Then, they used two approaches. The first is the creation of a text summary for the generation of the title (the title is the shortest summary of an article) in length limitation using a neural network with the encoder-decoder architecture (Transformer). The second is the generation of keywords ranked according to their relevance to the text. To do this they use an entity recognition service called off-the-shelf. And they use a ranking algorithm (XGBoost). In the end they combined these two approaches to generate the SEO title. So for the encoder they used a pre-trained German language BERT model to create a title generator for an article text, they modified the title generator by incorporating relevant keywords for SEO. And as the decoder is trained from scratch, they incorporated Fine tuning techniques (text summarization), as it optimizes the encoder and decoder. So the two important factors in their model are the length of the title and the keywords they contain [17].

The idea of Gjorgjevska, E., & Mirceva, G. in June 2021 is to perform automatic content categorization using the concept of semantic similarity between corpora. It's implementations of NLP and ML for better segmentation more profitable content to businesses on the website and for

more intelligently structured content. They proved that grouping web content into topics based on the semantic similarity between them gives better meaning to their performance outside of user search. Their process is as follows [18] :

- Define document corpus (download web page content with Scrapy) and result ranking using LDA.
- Convert multiple subjects to English.
- Analysis and pre-processing of data (tokenization, removal of stop words, special characters, HTML number tag, html, tags...etc.)
- Work with pre-trained word embeddings and calculate semantic similarity using Word2Vec for words and nBOW for documents.
- Create a matrix composed of values representing semantic similarity and perform hierarchical clustering.

LDA is used for ranking by topic as they encountered two issues. The first is the difficulty of finding the topic for a very large site. The second is that many companies do not organize their content on the web [18].

Polato, M., Demchenko, D., Kuanyshkereyev, A., & Navarino, N. in December 2021 create a model using deep learning for multilingual keyword classification. Because keywords are one of the information that deduce the interest of the user. Their model is based on the incorporation of keywords using neural networks such as FastText, because it allows to deal with the multilingual problem in an effective and efficient way and its architecture is inspired by the DeepSets model. They mentioned that FastText improves Word2Vec in several aspects, such as learning to represent internal word structures instead of the whole word, which gives FastText the ability to generate embeds for out-of-vocabulary (OOV) words, and the MLP is quick to train and efficient at the time of prediction. Their model is defined as follows [19]:

- Pre-processing (removal of special characters and lowercase the remaining characters).
- Split keyword into words and embed the words using FastText then feed the keyword into the model as a 2D (rows are the words and columns are the latent features of the words) setting the dimension to 10.
- The keyword representation is then fed into the DeepSets φ model network and the output is fed into a network (DeepSets ρ (the multilayer perceptron)) that performs the classification.

Table 3 examines each one of the works that improves SEO by mentioning the techniques used in the work, these benefits and their limits.

TABLE II. COMPARATIVE TABLE OF TECCHNIQUES THAT IMPROVE IR

Work	Used technics	Benefits	Limits
[10] in 2020	BERT	More precise than classical techniques. Embedding captures the semantics better.	Incorporating Big Data with nearest neighbor search is costly in terms of time complexity.
[13] in 2020	BERT KNN	The benefits of the previous work more the flexibility or user can set k and n according to his objectives.	The limit of the previous work in addition the beneficial depends on the performance of user.
[12] in 2021	LDA CombMNZ BAGG Ens ADA Ens	Competitive model and achieves good performance results.	LDA model loses its efficiency as the corpus grows.
[11] in 2022	BERT LDA	Better results than other techniques with lower loss training model.	Accuracy a little low because they have defined a few documents as if they have only one subject when they have several. Performance a bit low due to lack of data.
[14] in 2022	GPT BERT	High performance model with rich GPT-3 knowledge base with lower number of parameters.	The size of the data has an influence on the performance of the results.
[16] in 2019	BERT BM25	Performance increases according to the number of examples already seen during the training.	They compared BERT with other less efficient methods while neglecting the stronger techniques.
[15] in 2021	LSA BERT GPT Pegasus T5 FAISS	LSA is less expensive for calculating semantic similarity SBERT, DistilBERT gives better results than BERT or RoBERT.	BERT, GPT and T5 are very expensive and slow for calculating semantic similarity.

TABLE III. COMPARATIVE TABLE OF TECHNIQUES THAT IMPROVE IR

Work	Used technics	Benefits	Limits
[17] in 2021	BERT off-the-shelf XGBoost	The text summary is the best method to generate a title and the BERT classification of the keywords allows having excellent results.	Their decoder is not pre-trained and the performance of their model is strong for the German language.
[18] in 2021	Framework (Scrapy), NLP (LDA), ML, Word2Vec, nBOW	Powerful model improves e-reputation due to the combination of NLP and Deep learning techniques.	Extracting topics from a large site is difficult and poor web content architecture expose a problem.
[19] in 2021	FastText DeepSets MLP NB	Model achieved high accuracy scores FastText is fast and multilingual and it has the ability to generate integrations for OOV MLP is fast to train and efficient in prediction.	Limits on the form of keywords where each keyword cannot have more than 10 words. BERT is better in accuracy than the techniques using.

IV. RESEARCH ORIENTATIONS

According to our research that aims to improve e-reputation by improving IR and SEO. And according to our study of the different classic and modern AI techniques, we have found that the use of classical AI techniques is not sufficient to improve IR or SEO, because of their limitations such as weak capture of the semantics of similar contexts written with different words. However, they are used to compare documents in a large space [15], with the problems of polysemy and monosemy [12]. If we compare them, with new technologies we find that they have less efficient representations and the relevance depends on a very large dataset [4]. Even for a model of the new techniques, the collection of a large set of training data has a very strong influence on the performance of the model (it improves with the increase in data [11].

Although pre-trained neural models have achieved impressive results on different NLP spots [16]. The use of transformers such as BERT alone for the development of a system that improves either IR or SEO is very computationally expensive compared to other ranking algorithms, so their use is effective for reranking, spots on a small set of data [16], or to optimize a model (Finetuning) [13]. We need to test and study transformers more deeply, because each one is the best for a specific task. It seems that GPT-3 gives better results in terms of performance than BERT or BM25 even if the test data is not already seen during training due to its richer knowledge base [14].

According to [18] grouping into topics using the classic techniques of web content based on the semantic similarity between them gives better meaning of their performance and the use of modern techniques like BERT or GPT for ranking results gives strong relevance and performance of the system.

We propose in our future work to combine classical approaches based on keywords and modern approaches based on transformers such as BERT, GPT and T5 for the improvement of the semantic aspect of IR and SEO because it gives better results either for IR [11] [13] [15] or for SEO [18].

In addition to improving semantic information retrieval and semantic SEO, we plan in the future, to enrich it with closely related complements such as: topic modeling [35] [36], text classification [37], text summarization [38] [39] [40] similarity between texts [41] …etc.

V. CONCLUSION AND FUTUR WORK

From our research we assume that the combination of classical and modern AI techniques in the field of natural language processing (NLP) to improve IR and SEO allows to give better results because they minimize the error rate and give better accuracy and performance. Our objective is to continue our comparative study between the different AI techniques already studied and to use them to improve IR by integrating the semantic aspect into Big Data, whether in relevance or speed. In addition, improve the content stored in Big Data by integrating the semantic aspect for better referencing of its content. All these two researches in order to improve the e-reputation of the company.

REFERENCES

[1] Manning, Christopher D. "Introduction to information retrieval. Syngress Publishing," 2008.

[2] Zhang, Xinmeng and all. "Evolution analysis of Information Retrieval based on co-word network," 3rd international conference on Electronic Information Technology and Computer Engineering (EITCE). IEEE, 2019.

[3] Chunjiang, Cai. "E-commerce Search Engine Marketing Mechanism Analysis and Optimization," 13th International Conference on Intelligent Computation Technology and Automation (ICICTA). IEEE, 2020.

[4] Xu, Joyce. "Topic Modeling with LSA, PLSA, LDA & lda2Vec," NanoNets, on Medium, May 25 ,2018.

[5] Ketkar, Nikhil, and Eder Santana. "Deep learning with Python," vol. 1. Berkeley: Apress, 2017.

[6] Shah, Chintan, and Anjali Jivani. "A hybrid approach of text summarization using latent semantic analysis and deep learning," International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2018.

[7] https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/ consulted in 20/06/2022.

[8] Vaswani, Ashish, and all. "Attention is all you need," Advances in neural information processing systems 30, 2017.

[9] Ben Ammar, Anis. "Profiles in information retrieval: definition, exploitation and adaptation," Diss. Toulouse 3, 2003.

[10] Kaveti Naveenkumar, shrutendra harsola. "Deep Learning for Semantic Text Matching," Towards Data Science, 2020.

[11] Yang, Nakyeong, and all. "Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models," Expert Systems with Applications 190 , 2022.

[12] Ormeño, Pablo, Marcelo Mendoza, and Carlos Valle. "Topic Models Ensembles for AD-HOC Information Retrieval," Information 12.9, 2021.

[13] Rawal, Samarth. "Multi-Perspective Semantic Information Retrieval in the Biomedical Domain," Diss. Arizona State University, 2020.

[14] Kim, Su Young, and all. "Ask me what you need: Product Retrieval using knowledge from GPT-3," arXiv preprint arXiv:2207.02516 , 2022.

[15] Baronetzky, Pia, and all. "Deep Learning in Natural Language Processing for analysis of document similarity," 2021.

[16] Nogueira, Rodrigo, and Kyunghyun Cho. “Passage re-ranking with BERT,” arXiv preprint arXiv:1901.04085 , 2019.

[17] Anastasiu, Cristian and all. “DeepTitle--Leveraging BERT to generate Search Engine Optimized Headlines,” arXiv preprint arXiv:2107.10935, 2021.

[18] Gjorgjevska, Emilija and Georgina Mirceva. “Content Engineering for State-of-the-art SEO Digital Strategies by Using NLP and ML,” 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). IEEE, 2021.

[19] Polato, Mirko and all. “Efficient Multilingual Deep Learning Model for Keyword Categorization,” IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2021.

[20] Devlin, Jacob and all. “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[21] Radford, Alec and all. “Improving language understanding by generative pre-training,” 2018.

[22] Raffel, Colin and all. “Exploring the limits of transfer learning with a unified text-to-text transformer,” J. Mach. Learn. Res. 21.140 , 2020.

[23] Beliga, Slobodan, Ana Meštrović, and Sanda Martinčić-Ipšić. “An overview of graph-based keyword extraction methods and approaches,” Journal of information and organizational sciences 39.1, 2015.

[24] Merchant, Kaiz, and Yash Pande. “Nlp based latent semantic analysis for legal text summarization,” International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2018.

[25] Tu, H. T., Phan, T. T., & Nguyen, “An adaptive latent semantic analysis for text mining,” International Conference on System Science and Engineering (ICSSE) (pp. 588-593). IEEE, 2017.

[26] Hasan, HM Mahedi, Falguni Sanyal, and Dipankar Chaki. “A novel approach to extract important keywords from documents applying latent semantic analysis,” 10th International Conference on Knowledge and Smart Technology (KST). IEEE, 2018.

[27] Huaijin, Peng, Wang Jing, and Shen Qiwei. “Improving text models with latent feature vector representations,” IEEE 13th International Conference on Semantic Computing (ICSC), 2019.

[28] Huaijin, P., Jing, W., & Qiwei, S.“Improving text models with latent feature vector representations,” IEEE 13th International Conference on Semantic Computing (ICSC), pp. 154-157, 2019.

[29] Liang, Q., Wu, P., & Huang, C. “An efficient method for text classification task,” In Proceedings of the International Conference on Big Data Engineering, pp. 92-97, 2019.

[30] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” arXiv preprint arXiv:1406.1078, 2014.

[31] Caubriere, A. “From signal to concept: deep neural networks applied to speech understanding,” Doctoral dissertation, the Mans Université, 2021.

[32] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. B. “Pre-training of deep bidirectional transformers for language understanding In: Proceedings,” of the conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MN: Association for Computational Linguistics, 4171-86, 2019.

[33] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. “Language models are few-shot learners,” Advances in neural information processing systems, 33, 1877-1901, 2020.

[34] Mahapatra, Ajit Kumar, and Sitanath Biswas. “Inverted indexes: Types and techniques,” International Journal of Computer Science Issues (IJCSI) 8.4, 2011.

[35] CHIRAG GOYAL, “ Part 16 : Step by Step Guide to Master NLP - Topic Modelling using LSA,” 26 june 2021.

[36] MOHAMMED, Shaymaa H. et AL-AUGBY, Salam. “Lsa & lda topic modeling classification: Comparison study on e-books,” Indonesian Journal of Electrical Engineering and Computer Science, p. 353-362, 2020.

[37] ZHANG, Weiyu et XU, Can. “Microblog Text Classification System Based on TextCNN and LSA Model,” 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT). IEEE, p. 469-474, 2020.

[38] SHAH, Chintan et JIVANI, Anjali. “A hybrid approach of text summarization using latent semantic analysis and deep learning,” ,international conference on advances in computing, communications and informatics (ICACCI). IEEE, p. 2039-2044, 2018.

[39] ALLAHYARI, Mehdi, POURIYEH, Seyedamin, ASSEFI, Mehdi and all. “Text summarization techniques: a brief survey,” arXiv preprint arXiv:1707.02268, 2017.

[40] MA, Tinghuai, PAN, Qian, RONG, Huan and all. T-bertsum. “Topic-aware text summarization based on bert,” IEEE Transactions on Computational Social Systems, 2021.

[41] HUSSEIN, Ashraf S. “Arabic document similarity analysis using n-grams and singular value decomposition,” IEEE 9th international conference on research challenges in information science (RCIS). IEEE, p. 445-455, 2015.