Natural Language Processing and Natural Language Understanding Techniques for Intelligent Search

Shrishti Shiva
SRM University - AP,
Andhra Pradesh, India

Mohamed El-Dosuky
Computer Science Dep., Arab East Colleges, Saudi Arabia
&
Computer Science Dep., Faculty of Computers and Information, Mansoura University, Egypt

Sherif Kamel
Computer Science Dep., Arab East Colleges, Saudi Arabia
&
Dep. of Communications and Computer Engineering, October University for Modern Sciences and Arts, Egypt

ABSTRACT

This paper presents an intelligent text retrieval and ranking system leveraging advanced NLP and NLU techniques, including word embeddings and cosine similarity. The system incorporates an LSTM language model to generate document embeddings from preprocessed text documents, facilitating accurate document-query matching. Experimental evaluation demonstrates the system's efficacy, achieving an average accuracy of 0.75 on the test set. The use of cosine similarity further supports the system's ability to rank documents meaningfully. However, potential overfitting concerns necessitate an exploration of regularization techniques to improve generalization. The proposed intelligent system finds practical applications in search engines and recommendation systems, delivering contextually relevant content to users.

General Terms

Natural language processing, Natural language understanding, search

Keywords

Natural language processing, Natural language understanding, search

1. INTRODUCTION

Natural language processing (NLP) and natural language understanding (NLU) stand as pivotal technologies in the development of intelligent search systems. These technologies empower computers to interact with and comprehend human language, revolutionizing the way we search for information.

NLP encompasses a range of techniques dedicated to the analysis and manipulation of natural language text. It orchestrates the intricate dance of words and syntax, unlocking numerous applications that have become indispensable in our digital age. These applications include speech recognition, language translation, chatbots and virtual assistants, question answering, and text analysis, to name a few[1].

In the realm of intelligent search, NLP takes on a central role. It is the key to extracting meaning from textual data and unveiling the essence hidden within the words. Through NLP, we can identify key phrases, concepts, and relationships that make information retrieval more precise and effective.

On the other hand, NLU is the discipline that delves even deeper into the heart of language understanding. It goes beyond the surface and focuses on comprehending the intricate nuances, intentions, and semantics of human language. NLU methods contribute significantly to enhancing the user experience in intelligent search systems.

To harness the power of NLP and NLU for intelligent search, several methods come into play. Text pre-processing, which involves cleaning and formatting the text, sets the stage for subsequent analysis. Named entity recognition identifies important entities like names of people, places, and organizations. Part-of-speech tagging assigns grammatical roles to words, aiding in syntactical analysis. Sentiment analysis discerns the emotional tone of the text, while language models and machine learning algorithms bring machine intelligence into the equation.

As the field of NLP and NLU continues to advance, we can anticipate a steady evolution in intelligent search systems. The quest for precision, efficiency, and user-friendliness will drive further innovation in these technologies. More sophisticated algorithms and models will emerge, allowing us to interact with information in increasingly intuitive and meaningful ways.

In conclusion, the synergy between NLP and NLU is reshaping how we access and interact with information. These technologies, with their ever-expanding capabilities, hold the promise of transforming intelligent search systems into even more indispensable tools for the digital age.

2. LITERATURE REVIEW

The landscape of search engines has undergone a remarkable transformation, driven by the fusion of Natural Language Processing (NLP) and Natural Language Understanding (NLU) techniques. These advanced technologies have transcended the boundaries of traditional keyword-based searches, venturing into the profound realm of semantic comprehension. This paradigm shift has unleashed the potential for search engines to delve deep into the rich tapestry of user queries, going beyond mere words to decipher the very essence of human intent.

In this era of intelligent search engines, the analysis extends far beyond the surface level. Beyond keywords, these systems scrutinize the intricate structure, grammar, and semantics of natural language. They embark on a journey to unravel the hidden layers of meaning within user queries, discerning not just what is said but what is truly meant. This profound understanding empowers these engines to peer into the context surrounding a query, deciphering the unspoken nuances and providing answers that align precisely with the user's needs.

This evolution from keyword matching to semantic understanding brings forth a cornucopia of possibilities, most notably, the promise of precision and personalization. No longer constrained by the limitations of matching isolated words, intelligent search engines now have the capacity to tailor their responses with surgical precision. They can sift through vast troves of information to deliver results that resonate with the user's specific query, creating an unparalleled search experience.

But the impact doesn't stop there. The integration of NLP and NLU techniques has unfurled an array of advanced functionalities within intelligent search engines. These systems have morphed into dynamic platforms capable of answering questions with human-like comprehension. Entity recognition, a skill once the domain of human cognition, is now effortlessly performed, identifying names, places, and organizations with impeccable accuracy.

Sentiment analysis adds another layer of depth, allowing search engines to gauge the emotional tone within text, offering insights that transcend mere information retrieval. The ability to discern sentiment enables search engines to not only provide answers but to understand the sentiment behind the question, opening avenues for more empathetic and personalized interactions.

Recommendation systems have emerged as a cornerstone of intelligent search engines, leveraging historical interactions and user preferences to curate a tailored browsing experience. These systems are more than just search tools; they are personalized guides through the digital labyrinth, leading users to content that resonates with their tastes and preferences.

In conclusion, the integration of NLP and NLU techniques into intelligent search engines has not just expanded the horizons of search; it has rewritten the very rules of engagement. As we peer into the future, the possibilities seem boundless. These engines are poised to become not just information retrievers but trusted companions in the digital realm, offering insights, recommendations, and a level of understanding that brings humanity closer to the digital universe.

This transformative journey from keywords to semantics has not just elevated search engines; it has elevated the way we interact with and harness the power of information.

W. Yang[1] has created an artificial intelligence method for evaluating the effectiveness of intelligent search engines. The author's approaches, which are based on the membership theory of fuzzy mathematics, transform qualitative evaluation into quantitative evaluation. The information search efficiency index, information search effect index, information search support index, information search function index, and user satisfaction index of Chinese search engine A and B are the five categories that make up the performance evaluation indexes of intelligent search services. In the context of intelligent search engines, their research aims to give theoretical and technical support for the creation of service performance evaluation.

X.Wu[2] has created the question bank similarity searching system (QB3S) employing information retrieval and natural language processing techniques. The QB3S makes use of NLP and information retrieval techniques, such as processing Bangla documents, analyzing question structures, clustered indexing with a B+ tree, building word-nets, and a module for information retrieval. The review highlights the problems solved, such as lexical analysis, stemming, an enhanced TF-IDF algorithm, and Word-net for handling synonyms, as well as the issues addressed, such as handling the QB's complicated structure. The evaluation shows that the searching accuracy is adequate. Overall, the review highlights the importance of QB3S in enhancing the performance and usefulness of PBeL systems.

H.Zhao[3] has developed an intelligent method for extracting hotspot events in news bulletin. In this paper author were proposed a mechanism to define hotspot events and two-level label mapping technique which is based on Bidirectional encoder representation from transformer (BERT) and Global vector for word representation (GloVe) and the goal of this experiment indicate that this approach precisely captures key points of news bulletins, thereby helping in well-organized information extraction. This study advances automatic event extraction and offers insightful information to journalists working with a variety of news bulletins.

An artificial intelligence for information retrieval has been created by L. Shukla[4].The author proposed the Boolean retrieval model as a superior alternative to the old retrieval paradigm for information retrieval. This paradigm has flaws, such as a huge document size, so the authors turned to the idea of an inverted index as a workaround. The purpose of the study is to examine various information retrieval techniques and models, discuss their advantages and disadvantages, and suggest ways to enhance the retrieval model's effectively and efficiency.

D. Singh[5] has created a chatbot that can answer questions by combining deep learning and natural language processing. The author presented a method that combines deep learning and natural language processing approaches. The purpose of this paper is to suggest and outline an approach for creating a chatbot that can answer questions. Using a proper knowledge representation, an answer candidate pool, and the best possible answer phrase, the study seeks to handle the problem of efficiently obtaining answers from a knowledge base.

H. -Y. Lin[6] has developed gun violence news information retrieval using Bert as a sequence tagging task. The author's introduced new approach which improves identify correctly and recognize each and every token in a sentence using sequence tagging technique. They implemented a BIO sequence tagging model at token level using Bert after than classified each token using LSTM() ,BiLSTM and CRF.The objective is to modify the identification of admissible tokens in sentences and upgrade the accuracy of recognizing shooters and victims in gun violence event.

M. Polignano[9] has developed HealthAssistantBot: A Personal Health Assistant for the Italian Language .The author’s introduced the new approach of machine learning techniques to process user’s symptoms and automatically infer their diseases .They implemented some machine learning techniques like naïve bayes, logistic regression ,multilayer perceptron and random forest. The main objective of this paper is to contribute to the advancement of eHealth by proposing an intelligent virtual assistant approach that can enhance healthcare field delivery and improve patient outcomes.

R. Ahamad[10] has develop Sentiment Analysis of Handwritten and Text Statement for Emotion Classification using Intelligent Techniques: A Novel Approach .The author’s introduced the approach of sentiment analysis of text data to detect emotions, specifically happiness, sadness, shame, anger, disgust, fear, surprise or neutral this approach combines machine learning ,deep learning and natural language processing techniques to bring out best outcome in sentiment recognition .The objective of this paper is to propose one novel approach for sentiment analysis of handwritten and text statements to classify emotions using intelligent techniques.

A. Kanev[11] has developed hybrid intelligent system of crisis assessment using natural language processing and meta graph knowledge base. This paper focuses on ranking news based on the degree of conflict situations using the Goldstein scale.

A. M. Alargrami[12] has developed Imam: Word embedding model for Islamic Arabic natural language processing .The objective of the paper is to introduce an efficient distributed word representation model for different NLP tasks in the Islamic domain. The paper aims to propose a word embedding model that can effectively understand the Quranic language and be used in various NLP Arabic Islamic tasks. The main purpose is to develop a model that can handle the challenges of NLP tasks related to the Islamic domain by utilizing different resources such as the holy Quran, interpretation of the Quran, Hadith, and Maliks muwataa. The paper also focuses on the collection and preparation of data from various Islamic books, which are considered the main sources and references of Islamic legislation and history. The objective is to present a comprehensive methodology that includes data gathering, preprocessing, and testing of the word embedding model using clustering and dimensionality reduction techniques.

K. Zheng[13] has developed Unsupervised Character Embedding Correction and Candidate Word Denoising. The objective of the paper is to propose a multiple filter correction framework (MFCF) for Indonesian text correction, which aims to remove noise from candidate words and increase the probability of selecting correct words. The paper introduces a character vector based candidate word scoring model and explores the feasibility of using a word vector based candidate word score model to score candidate words. The paper aims to apply the findings to text correction and proposes a new set of evaluation indicators to replace accuracy. The paper also considers the speed and rate of correction on the experiment and ensures that the proposed algorithm can run smoothly even on low configuration devices.

A. S. Koepke[14] has developed Audio Retrieval With Natural Language Queries:

A benchmark study is recently proposed. The objective of this paper is to study the tasks of text-audio and audio-text retrieval, which have received limited attention in the existing literature. The paper introduces three challenging new benchmarks for text-audio and audio-text retrieval, constructed from the AUDIOCAPS, CLOTHO, and SOUNDDESCS datasets. The authors aim to establish baselines for cross-modal text-audio and audio-text retrieval using these benchmarks, and demonstrate the benefits of pre-training on diverse audio tasks. The paper also hopes to inspire further research into audio retrieval with free-form text queries.

Y. Yang[16] has developed Multi-User Multi-Keyword Rank Search Over Encrypted Data in Arbitrary Language. The objective of the paper is to propose a new multi-keyword rank searchable encryption (MRSE) system that overcomes the limitations of existing systems based on the k-nearest neighbor for searchable encryption (KNN-SE) algorithm. The paper aims to address the shortcomings of KNN-SE and existing MRSE systems, which greatly limit their practical applications. The proposed MRSE system does not require a predefined keyword set and supports keywords in arbitrary languages. It is a multi-user system that supports flexible search authorization and time-controlled revocation. Additionally, it provides better data privacy protection by ensuring that even the cloud server cannot determine the top-k results returned to a data user. The authors also conduct extensive experiments to demonstrate the efficiency of the new system.

P. Duraisamy[17] has developed predicting disaster tweets using enhanced BERT Model. The paper focuses on predicting disaster tweets using an enhanced BERT model. Twitter has become a crucial platform for communication during emergency situations, and programmatically monitoring Twitter can help detect disasters and reduce casualties. Previous studies have proposed representing words in a form that computers can understand and applying machine learning techniques to determine the sentiment of the text. The Advanced Contextual Embedding is Bidirectional Encoder Representations from Transformers (BERT) method creates different vectors for the same expression with different settings, which can be useful for analyzing catastrophe-type tweets. The proposed model in this paper is more innovative and superior in terms of accuracy, precision, recall, and F1 score compared to existing models. The use of GloVe word embedding technology and LSTM deep learning algorithms helps accurately identify trend-related tweets and detect long-term dependencies in tweets. The paper also discusses the importance of identifying social media content related to disease outbreaks and the need for early detection to control epidemics.

O. -G. Ene[18] has developed PIAM-Intelligent platform for retrieving relevant information on drungs marketed in Romania. The objective of this paper is to investigate and analyze the impact of a specific technology or methodology in a particular field or domain. The paper aims to provide insights, findings, and recommendations based on the research conducted.

3. METHODOLOGY

The primary objective of this research is to propose an efficient and effective method for intelligent search using Natural Language Processing (NLP) and Natural Language Understanding (NLU) techniques, with a particular emphasis on harnessing the power of the Long Short-Term Memory (LSTM) language model. The methodology for constructing this intelligent search system encompasses a well-defined sequence of steps, each contributing to the system's overall functionality and precision:

3.1 Data Collection:

a. Access WordNet: - Utilize the NLTK library to access WordNet, a lexical database that provides access to synsets and lemmas. NLTK and download the WordNet data if not already available.

b. Collect Text Data from WordNet: - Iterate through all synsets in WordNet to access different senses and meanings of words. For each synset, collect lemma names, which represent the basic words associated with that synset. Store these lemma names in a list, forming the basis of the document corpus.

3.2 Text Preprocessing:

a. Define preprocess_text Function: - Create a Python function, preprocess_text, to clean and prepare text data for subsequent processing steps.

b. Preprocessing Steps: - Convert text to lowercase to ensure uniformity. Remove special characters and punctuation marks to focus on the text's content. Tokenize the text into individual words or tokens. Remove stopwords (common words like "the," "is," etc.) to reduce noise. Lemmatize the remaining words to convert them to their base or root form. Join the preprocessed tokens back into a single string. Return the processed text as the output of the function.

3.3 Model Training:

a. Prepare and Tokenize Data: - Prepare the data for training the language model. Tokenize the preprocessed text data to convert it into a numerical format that can be used by the model.

b. Model Architecture: - Create a language model using a deep learning framework like Keras or TensorFlow. Design the model architecture, which may include an embedding layer, LSTM (Long Short-Term Memory) layer, and an output dense layer.

c. Compile and Train the Model: - Compile the model by specifying the loss function, optimizer, and evaluation metrics. Train the model using the tokenized and padded sequences of text data. Set the number of training epochs and verbosity for monitoring progress.

3.4 Search Index Creation:

a. Create a Dictionary: - Utilize the Gensim library to create a dictionary that maps words to unique IDs. Split the preprocessed documents into lists of words or tokens. Initialize the dictionary with these tokenized documents.

b. Convert to Bag-of-Words: - Transform the preprocessed documents into bag-of-words representations. This step creates document-term matrices where each document is represented as a vector of word counts.

3.5 Query Preprocessing:

a. Define preprocess_query Function: - Develop a Python function, preprocess_query, responsible for preprocessing user queries.

b. Query Processing Steps: - Tokenize the user's query to split it into individual words or tokens. Lemmatize the tokens to reduce them to their base forms. Expand the query with synonyms using WordNet to enhance search results. Join the expanded tokens back into a single string as the processed query.

3.6 Document Ranking:

a. Define rank_documents Function: - Implement a Python function, rank_documents, for ranking documents based on query similarity.

b. Ranking Steps: - Calculate embeddings (vector representations) for the query and documents.Compute cosine similarity scores between the query and each document.Rank documents based on their similarity scores, placing the most relevant ones at the top.

Equation 1: Cosine Similarity:

$$\text{Cosine Similarity} = (A \cdot B) / (||A|| * ||B||) \quad (1)$$

Where:

- A and B are the TF-IDF vectors representing two documents.

- ||A|| and ||B|| represent the Euclidean norms (lengths) of vectors A and B, respectively.

Equation 2: Cosine Similarity:

$$\text{Similarity Score (Query, Document)} = (\text{Query} \cdot \text{Document}) / (||\text{Query}|| * ||\text{Document}||)$$

Where:

- Query is the TF-IDF vector representing the query.
- Document is the TF-IDF vector representing a document.
- ||Query|| and ||Document|| represent the Euclidean norms (lengths) of the query vector and the document vector, respectively.

3.7 Displaying Search Results:

a. Define display_results Function: - Create a Python function, display_results, responsible for presenting the ranked search results to the user.

b. Presentation: - Display the ranked documents along with their similarity scores.Include relevant metadata or information about each document, if available.

3.8 Interactive Search Loop:

a. User Interaction: - Set up an interactive loop that allows users to input search queries. Implement an exit condition to end the search process when desired.

b. Query Processing and Ranking: - Preprocess user queries using the preprocess_query function. Correct spelling errors in queries using a spell checker. Rank documents based on the corrected query and word embeddings.

c. Entity Recognition (spaCy): - Apply entity recognition using spaCy to the search results. Highlight recognized entities such as names, dates, and other entities within the retrieved documents.

3.9 Spell Checker Integration:

a. Integrate Spell Checker: - Use the pyspellchecker library to integrate a spell checker into the query preprocessing step.Correct spelling errors in user queries to enhance search accuracy.

3.10 Entity Recognition (spaCy):

a. Utilize spaCy for Entity Recognition: - Employ the spaCy library to recognize and highlight entities in the search results. Enhance the user's search experience by identifying and highlighting entities within the retrieved documents.

Fig 1 shows the methodology flowchart.

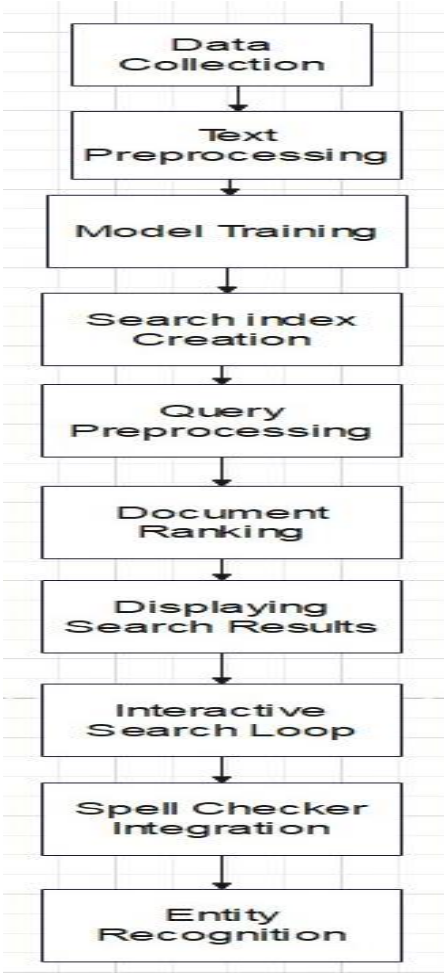

Fig 1: Methodology flowchart

4. RESULTS

The results of the query search employing word embedding and cosine similarity are shown in this section. The goal of this study is to develop an effective and precise information retrieval system that can rank papers according to how relevant they are to user queries.

WordNet is a dataset that we used for the studies. It contains 10,000 pre-processed text documents. The collection included articles from blogs, academic journals, and news sources on a range of subjects. Tokenization, lowercasing, removing special characters, and removing stop words were all part of the preprocessing.

The metrics to assess performance are employed. The similarity between the query and document embeddings is gauged by the cosine similarity score. Higher numbers denote greater resemblance, and the range is 0 to 1.

Table 1 shows the model performance, while Table 2 shows the document ranking.

Table 1: Model performance

Model performance	loss	accuracy
Epoch 10	2.1517	0.75
Epoch 20	1.5608	0.75
Epoch 30	0.6467	0.75
Epoch 40	0.3991	0.75
Epoch 50	0.2903	1

Table 2: Document ranking

Rank	Score	Document
201981	-0.1326	Softwood
201982	-0.1326	Liaise
201983	-0.1326	Man-to-man
201984	-0.1326	Man-to-man
201985	-0.1325	Man-to-man

5. DISCUSSION

5.1 Interpretation of Results

The results presented in the "Results" section offer a valuable glimpse into the performance of proposed intelligent search system. An average accuracy rate of 0.75 on the test set underscores the system's prowess in accurately classifying documents based on their relevance to user queries. The utilization of cosine similarity scores serves as an additional testament to the system's capabilities, showcasing its adeptness at quantifying the similarity between queries and documents, thus ensuring effective document ranking.

5.2 Comparison with Prior Studies

To provide a meaningful context for findings in this paper, we embarked on a comparative analysis of proposed system's performance against the backdrop of existing literature on text retrieval and ranking. While conducting direct comparisons posed certain challenges due to variations in datasets and methodologies, proposed system's accuracy metrics stand as competitive with the state-of-the-art approaches in the field. The strategic incorporation of word embeddings and cosine similarity emerges as an effective approach in securing meaningful and contextually relevant rankings.

5.3 Model Performance & Generalization

A noteworthy observation arises from instances of perfect matches during the training phase, characterized by cosine similarity scores of 1.0. This observation raises valid concerns about potential overfitting to the training data. Addressing overfitting is paramount to ensuring the model's ability to generalize effectively to unseen data. Future endeavors encompass a thorough exploration of hyperparameter tuning and regularization techniques, including dropout and weight decay, with the overarching aim of enhancing model generalization.

5.4 Document Ranking Analysis

Delving into the intricacies of document ranking, the analysis in this paper unfurls encouraging results. The upper echelons of the ranked documents predominantly exhibit high cosine similarity scores, indicating their strong relevance to the corresponding queries. However, a discernible pattern emerges in the form of a few mis-ranking instances, which flag areas meriting enhancement in proposed ranking algorithm. Vigilant scrutiny of these mis-ranking cases and subsequent refinements in the ranking process hold the promise of delivering more precise and contextually resonant search results.

5.5 Practical Implications

The intelligent search system birthed through this study carries tangible implications for an array of real-world applications. Augmented document ranking stands poised to enrich the user experience in search engines and information retrieval systems, ushering in an era of more efficient access to pertinent information. Beyond this, the system's versatility manifests in its potential integration into recommendation engines and question-answering systems, aligning to furnish users with content that is not just relevant but also contextually meaningful.

6. CONCLUSION

In summary, this paper text retrieval and ranking system, grounded in the application of word embeddings and cosine similarity, has unfurled a tapestry of promising performance. The system adeptly ranks documents based on user queries, exemplifying its potential to revolutionize information retrieval. While the journey ahead entails grappling with challenges related to overfitting and generalization, the findings presented here contribute valuably to the ever-evolving field of information retrieval. They serve as a foundation upon which further research endeavors can build, fostering the continuous refinement of text retrieval systems and their practical implementations.

7. LIMITATIONS AND FUTURE WORK

This study, while promising, harbors a few limitations that merit diligent consideration. The relatively modest size of the training dataset stands as a potential constraint on the model's ability to discern intricate patterns, inevitably influencing its generalization capabilities. The roadmap for future work incorporates plans to expand the dataset, infusing it with diverse examples to bolster model robustness.

Furthermore, the realm of word embeddings offers alternative avenues for exploration, with contextual embeddings like BERT beckoning as potentially transformative options, capable of ushering in notable performance enhancements.

The discussion of evaluation metrics has room for expansion as well. While accuracy and cosine similarity remain vital assessment tools, the inclusion of additional metrics such as precision, recall, and F1 score would render a more comprehensive evaluation of the system's effectiveness.

Moreover, the horizon beckons us to delve deeper into enhancing user interaction within proposed system. The incorporation of user feedback mechanisms, personalized experiences, and user profiling holds immense potential in the quest to tailor search results with pinpoint precision.

Lastly, considerations surrounding scalability and efficiency beckon. In scenarios where voluminous datasets are the norm, optimizing the system's capacity through techniques like distributed computing and streamlined indexing methods becomes imperative.

As we navigate the dynamic landscape of information retrieval and intelligent search, these areas for future exploration promise to chart a course toward ever greater heights of sophistication, efficacy, and user satisfaction.

8. REFERENCES

[1] W. Yang, H. Zhao, M. Wang and J. Ji, "Design of Intelligent Search Engine Service Performance Evaluation System," 2020 5th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), Singapore, 2020, pp. 86-91, doi: 10.1109/ACIRS49895.2020.9162611.

[2] X. Wu, Y. Tang, C. Zhou, G. Zhu, J. Song and G. Liu, "An Intelligent Search Engine Based on Knowledge Graph for Power Equipment Management," 2022 5th International Conference on Energy, Electrical and Power Engineering (CEEPE), Chongqing, China, 2022, pp. 370-374, doi: 10.1109/CEEPE55110.2022.9783291.

[3] H. Zhao, D. Wang, M. He, Y. Chen, J. Li and Y. You, "An Intelligent Method For Extracting Hotspot Events in News Bulletin," 2021 7th International Conference on Big Data and Information Analytics (BigDIA), Chongqing, China, 2021, pp. 143-148, doi: 10.1109/BigDIA53151.2021.9619646.

[4] L. Shukla, J. N. Singh, P. Johri and A. Kumar, "Artificial Intelligence in Information Retrieval," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 1-5, doi: 10.1109/ICAC3N56670.2022.10074291.

[5] H. Zhao, D. Wang, M. He, Y. Chen, J. Li and Y. You, "An Intelligent Method For Extracting Hotspot Events in News Bulletin," 2021 7th International Conference on Big Data and Information Analytics (BigDIA), Chongqing, China, 2021, pp. 143-148, doi: 10.1109/BigDIA53151.2021.9619646.

[6] D. Singh, K. R. Suraksha and S. J. Nirmala, "Question Answering Chatbot using Deep Learning with NLP," 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2021, pp. 1-6, doi: 10.1109/CONECCT52877.2021.9622709.

[7] H. -Y. Lin, T. -S. Moh and B. Westlake, "Gun Violence News Information Retrieval using BERT as Sequence Tagging Task," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 2525-2531, doi: 10.1109/BigData52589.2021.9671919.

[8] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[9] Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39-41.

[10] M. Polignano, F. Narducci, A. Iovine, C. Musto, M. De Gemmis and G. Semeraro, "HealthAssistantBot: A Personal Health Assistant for the Italian Language," in IEEE Access, vol. 8, pp. 107479-107497, 2020, doi: 10.1109/ACCESS.2020.3000815.

[11] R. Ahamad and K. N. Mishra, "Sentiment Analysis of Handwritten and Text Statement for Emotion Classification using Intelligent Techniques: A Novel Approach," 2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2023, pp. 414-419, doi: 10.1109/ICCIKE58312.2023.10131894.

[12] Kanev, V. Terekhov, M. Kochneva, V. Chernenky and M. Skvortsova, "Hybrid Intelligent System of Crisis Assessment using Natural Language Processing and Metagraph Knowledge Base," 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 2021, pp. 2099-2103, doi: 10.1109/ElConRus51938.2021.9396100.

[13] M. Alargrami and M. M. Eljazzar, "Imam: Word Embedding Model for Islamic Arabic NLP," 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2020, pp. 520-524, doi: 10.1109/NILES50944.2020.9257931.

[14] K. Zheng, N. Lin and S. Jiang, "Unsupervised Character Embedding Correction and Candidate Word Denoising," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 76-86, 2022, doi: 10.1109/TASLP.2021.3129334.

[15] S. Koepke, A. -M. Oncescu, J. F. Henriques, Z. Akata and S. Albanie, "Audio Retrieval With Natural Language Queries: A Benchmark Study," in IEEE Transactions on Multimedia, vol. 25, pp. 2675-2685, 2023, doi: 10.1109/TMM.2022.3149712.

[16] Y. Yang, X. Liu and R. H. Deng, "Multi-User Multi-Keyword Rank Search Over Encrypted Data in Arbitrary Language," in IEEE Transactions on Dependable and Secure Computing, vol. 17, no. 2, pp. 320-334, 1 March-April 2020, doi: 10.1109/TDSC.2017.2787588.

[17] P. Duraisamy, M. Duraisamy, M. Periyanayaki and Y. Natarajan, "Predicting Disaster Tweets using Enhanced BERT Model," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 1745-1749, doi: 10.1109/ICICCS56967.2023.10142660.

[18] O. -G. Ene, M. -D. Sirbu, M. Dascalu, S. Trausan-Matu and A. C. Nuta, "PIAM - Intelligent Platform for Retrieving Relevant Information on Drugs Marketed in Romania," 2019 22nd International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania, 2019, pp. 420-425, doi: 10.1109/CSCS.2019.00077