IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Question Answering System: An Nlp Based Approach

[1] Selvakumar G, [2] Thilaheswaran KK, [3] Devadarshan VG

[1] Assistant professor (Sr. G)

[1] Department of Artificial Intelligence and Data Science,

KPR Institute of Engineering and Technology, Coimbatore, India.

Abstract: To make the most of the abundance data available, effective tools like question answering (QA) systems are essential in this digital age, where knowledge is both plentiful and elusive. A domain-specific QA system designed for the dynamic field of computer science is proposed in the research. By utilizing Natural Language Processing and Information Retrieval (IR) methods, the system seeks to provide intelligent and helpful responses to users' questions at critical moments. It helps to improve the response; the research presents finding a perfect answer for a question using summarization. To generate a wide range of natural language responses, large language models will be used, with context-aware techniques guiding the best selection process using deep learning. The computer science community's ability to acquire knowledge and make decisions might be completely transformed by the QA system, which would eventually spur innovation and advancement.

Keywords: Information Retrieval (IR), Question Answering (QA), Natural Language Processing.

I. Introduction

In this era of information overload, where piles of data are always calling, it can seem impossible to separate the specific, trustworthy answers from the noise. Question answering (QA) system helps in finding answers to questions raised by humans in an efficient manner. QA system utilizes Natural Language Processing (NLP), and Information Retrieval (IR) to develop a find answer for the given question [1]. Compared to keyword search engines, they provide consumers with a more natural and user-friendly manner to access only the information. The users would be able to get the needed information in a quicker and precise way thus saving time for everyone in researching and figuring out better solutions. Question-answering systems are radically altering knowledge acquisition and comprehension, opening new avenues for study, research, and day-to-day activities.

The proposed QA system specializes in answering questions based on Computer Science. Programmers, developers, and researchers can utilize it and benefit from the QA system. In the dynamic field of computer science, where knowledge is the engine of growth and innovation is the norm, the pursuit of effective information access never goes away. QA systems would try to provide on-point answers rather than providing multiple paragraphs of content related to the question or the topic [2]. The system will be domain specific, and the proposed paper will focus on questions related to the computer science field. So that the developers can swiftly find answers to the questions in an easy way.

The demand for website scraping arose from the popularity of the Internet and the resulting rapid increase in the volume of data created every day. In simple terms, web scraping provides this feature in computer applications that can complete it more accurately and much faster than a human, as opposed to copying and pasting information manually from the internet and compiling it in a spreadsheet [3]. The process of turning unstructured web data into organized data that can be stored and examined in a single database or excel is known as web scraping. Here, the data is collected using web scraping.

An important task in QA systems is question classification as the type of question defines what type of answer we require. By narrowing the search field of possible responses, Question Classification enhances QA by facilitating better precision and effective answer discovery [4]. A key component of efficient response creation in the dynamic world of Q&A systems is the capacity to recognize a question's level of complexity. It explores the intricacies of language, the range of information necessary, and the solution necessary to get at a truly rewarding response. The question is then preprocessed using multiple processes like Latent Direct Allocation,

Next step would be generating the answer for the question and the Large Language Models (LLMs) have become maestros of language generation, churning out diverse and often human-quality responses to our queries. Yet, choosing the "best" response from this symphony of possibilities becomes a critical challenge. So, a summarization method is created to provide the best possible answer for a question is selected from a set of answers. Traditional evaluation methods, like random selection or simple fluency measures, fall short, leaving hidden gems of insightful and contextually relevant answers undiscovered.

II. Literature Survey

Latent Dirichlet Allocation (LDA) has been used along with Machine learning models like Naïve Bayes and Random Forest for tag classification thus achieving 90 percent recall for individual tag prediction [5]. As the tags play an important role in QA system helping to understand what kind of question it is. A proper ranking function of documents along with a proper preprocessing technique using vectorization provide QA systems with better precision [6]. The model shows us that with improvised ranking system it also reduces time of response. Utilizing transformer models helps in the response creation as well as question classification when trained with large amount of data beforehand, it would be easier to finetune the model. Using BERT and RoBERTa was used for customer health question classification and gave promising results with nearly 80 percent accuracy rates [7]. There will be millions of questions related to a single topic. For example, in stack overflow, there are 60,028 machine learning related questions [12].

Table 1 Literature Survey

Author	**Title & Year**	**Methodology**
Kavuk, E. M., & Tosun, A.	Predicting Stack Overflow question tags: a multi-class, multi-label classification (2020).	LDA for feature extraction from posts, utilizes one-against-all classifiers for 15 popular tags, and employs a multi-tag classifier to recommend the top K tags for each post.
T.N. Manjunath a, Deepa Yogish b, S. Mahalakshmi, H.K. Yogish d	Smart question answering system using vectorization approach and statistical scoring method (2020).	The smart question answering system employs Natural Language Processing, normalized term weighting (TF-IDF) with cosine similarity, and an enhanced BM25 ranking function for efficient document retrieval, achieving high precision (93.2%), recall (84.3%), and accuracy (80%).

Maliha Noushin Raida, Zannatun Naim Sristy, Nawshin Ulfat, Sheikh Moonwara Anjum Monisha, Md. Jubair Ibna Mostafa, Md. Nazmul Haque	Representation-centric approach for classification of Consumer Health Questions (2023).	Utilizing diverse representation techniques, achieves high accuracy in Consumer Health Question classification with fine-tuned BERT-large (86.00% GARD, 80.40% Yahoo! Answers) and RoBERTa-large (86.20% GARD) models, surpassing previous methods
Wai Keat Lee,Moon Ting SU	Mining Stack Overflow for API class recommendation using DOC2VEC and LDA (2021) [13]	Developed a recommendation approach using Doc2Vec for NL query-to-API documentation gaps, employing LDA for API class extraction from Stack Overflow, yielding a simplified yet effective solution.
N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas and G. Kalliris, A.	Modeling stack overflow tags and topics as a hierarchy of concepts (2019).	Treat each Stack Overflow question as a document, focusing on tag-driven relationships using the Label-to-Hierarchy model.
Maliha Noushin Raida, Zannatun Naim Sristy, Nawshin Ulfat, Sheikh Moonwara Anjum Monisha, Md. Jubair Ibna Mostafa, Md. Nazmul Haque	A study on classifying Stack Overflow questions based on difficulty by utilizing contextual features (2024).	Integration of TF-IDF, LDA, and Doc2Vec models with diverse classifiers and additional features to measure question difficulty, validated on three datasets, highlighting Doc2Vec's superior performance.
Tóth, László, Balázs Nagy, Dávid Janthó, László Vidács, and Tibor Gyimóthy.	Towards an accurate prediction of the question quality on stack overflow using a deep-learning-based nlp approach (2019) [14].	Utilized natural language processing and deep learning with Keras and TensorFlow to classify Stack Overflow questions based on quality, extracting textual information, and employing document vectors.
Cabrera-Diego, L. A., Bessis, N., & Korkontzelos, I.	Classifying emotions in Stack Overflow and JIRA using a multi-label approach (2020).	Employs HOMER and RAkEL multi-label classifiers enriched with features from various lexica to analyse emotions in Open-Source Software-related text from Stack Overflow and JIRA Issue Tracker.

Hassan, Sk Adnan, Dipto Das, Anindya Iqbal, Amiangshu Bosu, Rifat Shahriyar, and Toufique Ahmed.	Integrating question classification and deep learning for improved answer selection (2018).	Annotated programming questions on Stack Overflow were labeled for difficulty, the dataset was preprocessed, and ten supervised learning algorithms were applied using 10-Fold cross-validation for empirical validation.
Savenkov, Denis, and Eugene Agichtein.	CRQA: crowd-powered real-time automatic question answering system (2016) [15].	CRQA employs an automatic question answering module using Information Retrieval, generating candidate answers from diverse sources. The crowdsourcing module seeks worker-generated answers and ratings, with a re-ranking model aggregating this feedback for final answer selection.

III. METHODOLOGY

The methodology and architecture of the system proposed will be discussed in detail with the following sections.

DATASET:

The Stack Overflow (SO) and other question-and-answer sites are popular among developers as places to discuss and work through technical problems. Consequently, there are almost 16,000,000 unique queries on SO that address issues faced by developers. There are more than 27,000,000 response answers for each of these queries [16]. The stack overflow questions can be collected using an API known as Stack Exchange, utilizing this API questions and their answers were scrapped from the SO.

The dataset developed had 10,000 rows and two features. After looking through the dataset more thoroughly, the duplicate records that were created twice were eliminated because they had more than one tag tags [17].

PREPROCESSING:

Before extracting relevant information from the questions, preprocessing is necessary. The preprocessing processes were applied to the collected dataset should be of a preferrable one to make the modelling better. Common English terms or stop words along with punctuation marks, digits, excess white spaces, emojis and code snippets were eliminated along with any hyperlinks [18]. Next, every character is converted to lowercase and rearranged words so that they matched their root, a process known as stemming. For this, the Porter Stemmer from NLTK library was utilized. Thus, now the preprocessing is done, next would be feature extraction.

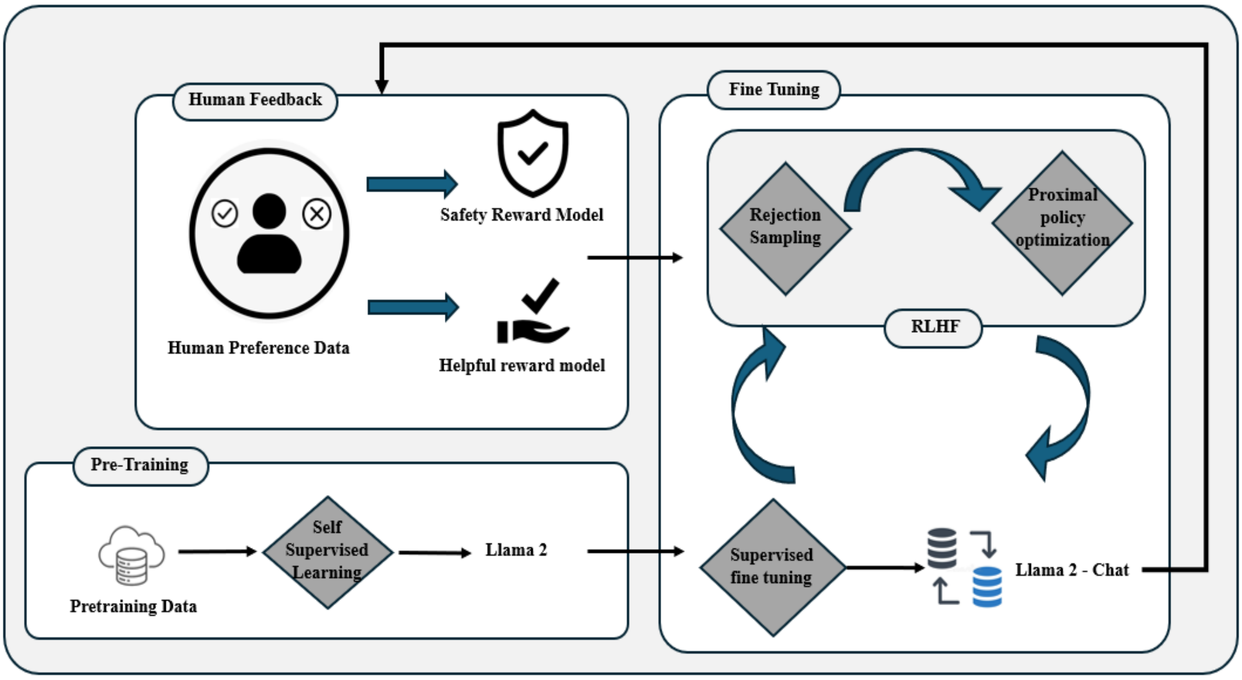


Figure 1 Workflow of the Model

MODELS:

In NLP, transfer learning is being used quite frequently nowadays. Sequential fine-tuning is the method of transfer learning that is most frequently used. Many textual-based natural language processing (NLP) issues have been resolved with the use of such trained language models for knowledge communication [19]. Strong performance on numerous NLP tasks is achieved through transfer from pre-trained models. Let us look into some of the pre-trained models and how we utilize it in the proposed paper.

BERT:

BERT is Bidirectional Encoder Representations from Transformers, a pre-trained model suited mainly for NLP tasks as training the dataset with pre-trained models, it enhances the answering ability of the QA system.

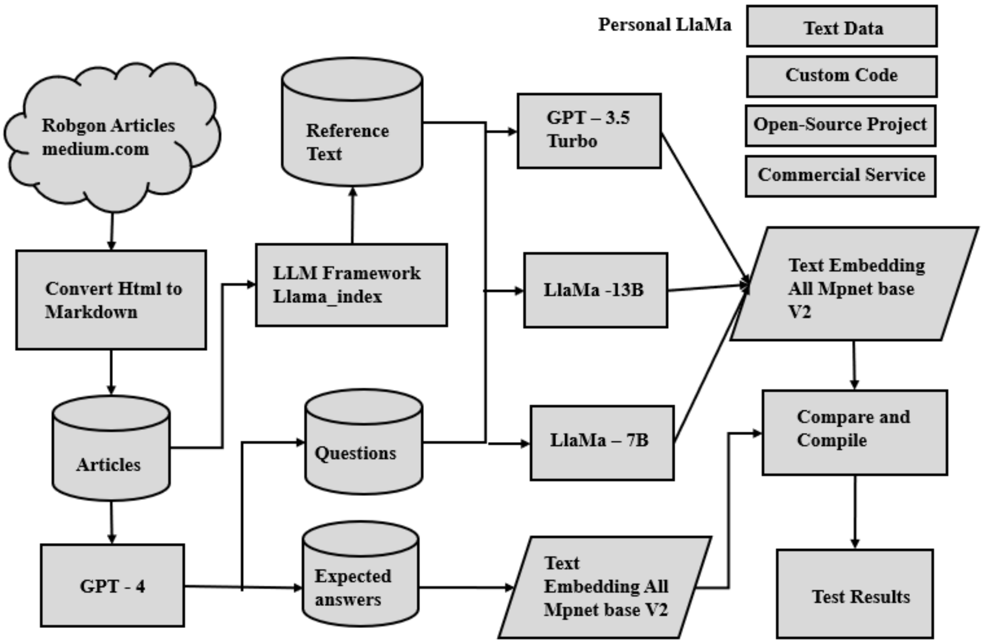


Figure 2 Architecture of BERT used in Model.

BERT is simply adjustable for a wide range of downstream operations. This is mostly attributable to the Transformer's self-attention mechanism, which enables BERT to promote generalizability.
BERT's lack of a clear concept of word order beyond simply labelling every single word with its exact embedding is another intriguing feature [20]. The BERT model forgoes recurrence structures and is instead a stack of multi-layer Transformer blocks. It draws global dependencies between input sequences only by using the self-att Llama 2:

Llama 2 model was released in 2023 by META AI and the main objective to this large language model was to improve NLP tasks like content generation, conversational AI, language translation. It has multilingual capabilities as well as enhanced scalability [22]. Among the most recent technologies is the Llama 2 model is noteworthy, especially for generative AI applications in computing solutions. Using a fresh blend of publicly accessible data, Llama 2, an upgraded version of Llama 1, was trained. Additionally, doubled the model's context length, added 40% to the pretraining corpus, and implemented grouped-query attention [23].

ention technique [21]. The transformer blocks, self-attention heads count in a transformer block and the hidden size, are the three architecture parameter settings for the BERT model. With a vocabulary of 30522 words, the pre-trained model employs the publicly available BERT base model (12 layers, 768 hidden dimensions, and 12 attention heads). The dropout probability in between transformer layers is set to 0.1.

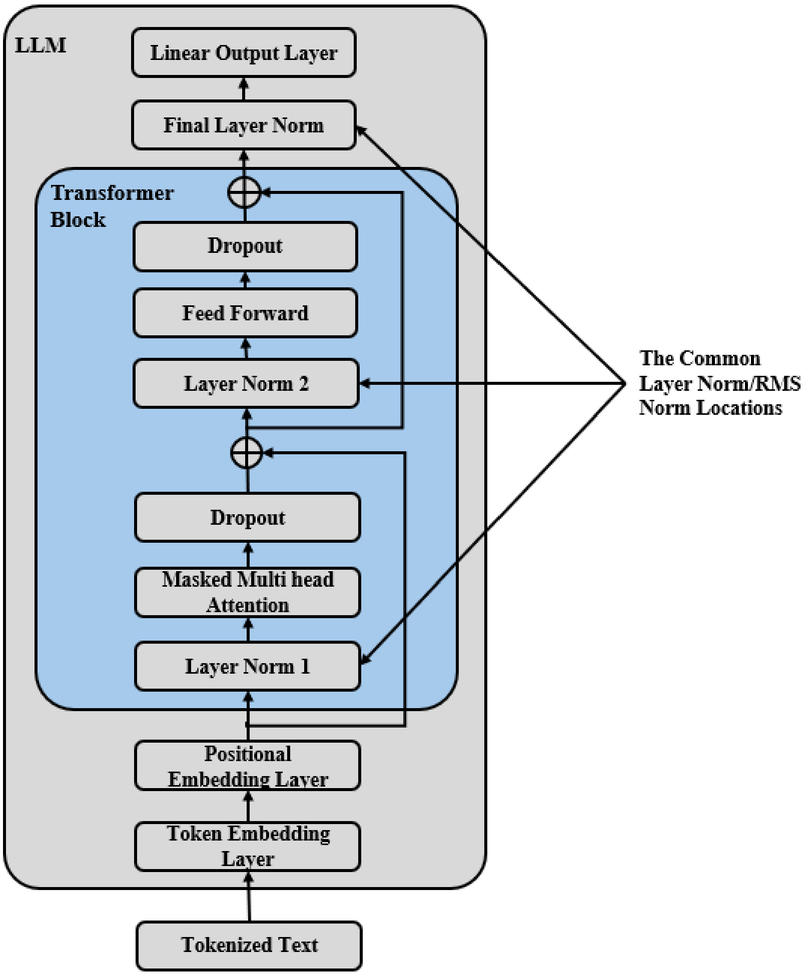


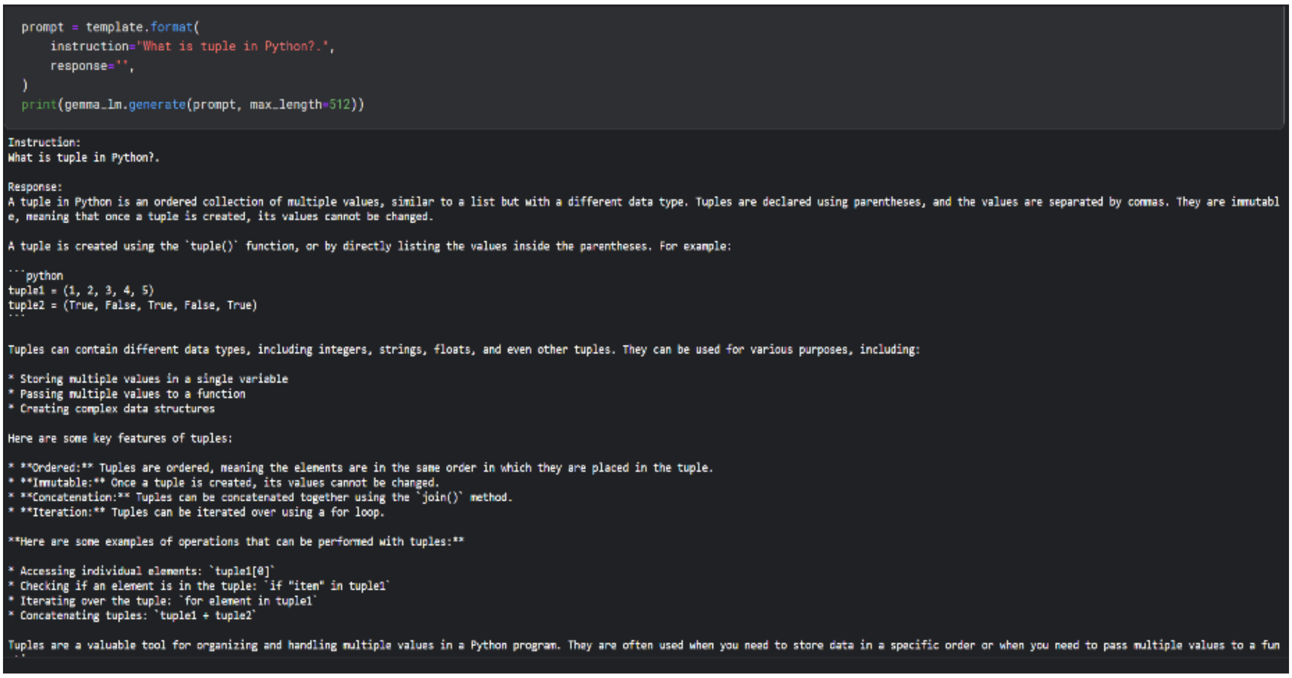
Figure 3 Architecture of Llama models

IV. Conclusion and Future Works

The proposed QA system presents a promising solution for streamlining information access within the computer science domain. By seamlessly integrating NLP, answer summarization, and LLMs, the system strives to provide users with accurate and tailored responses, saving valuable time and effort. This not only fosters informed decision-making but also fuels innovation in the ever-evolving world of computer science. While the proposed QA system presents a promising solution for computer science related queries, the journey doesn't end here. To ensure its real-world success, future research must address key challenges. Maintaining high-quality, unbiased data and fostering user trust through explainable reasoning are crucial for reliable and ethical outcomes. Personalization, adaptability, and integration with existing tools can further enhance the system's relevance and seamlessness. Rigorous evaluation using benchmarks and real-world studies will be essential to measure its impact and identify areas for improvement. Beyond its immediate benefits, this system holds the potential to democratize knowledge, personalize learning, and empower individuals at all levels of expertise. Addressing ethical considerations like data privacy and responsible information retrieval are paramount as we navigate the future of AI-powered knowledge access.

V. Results

Our evaluation of Llama, Gemma, and BERT models for automatic answering systems revealed high accuracy rates across predefined datasets. BERT exhibited slightly superior performance, particularly in nuanced and complex questions, while Llama and Gemma showed competitive accuracy on simpler queries. Robustness analyses confirmed the models' resilience against noise and adversarial inputs, with all maintaining high accuracy levels. Moreover, efficiency assessments indicated Llama and Gemma's comparable performance with lower computational demands compared to BERT. Generalization tests showcased their ability to answer questions from diverse sources effectively. These findings suggest the potential of these models in enhancing natural language processing tasks, with implications for future optimizations and integrations for improved performance in real-world applications.



REFERENCES

[1] Allam, Ali Mohamed Nabil, and Mohamed Hassan Haggag. "The question answering systems: A survey." *International Journal of Research and Reviews in Information Sciences (IJRRIS)* 2, no. 3 (2012).

[2] Mervin, R. "An overview of question answering system." *International Journal Of Research In Advance Technology In Engineering (IJRATE)* 1 (2013): 11-14.

[3] Khder, Moaiad Ahmad. "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application." *International Journal of Advances in Soft Computing & Its Applications* 13, no. 3 (2021).

[4] Madabushi, Harish Tayyar, Mark Lee, and John Barnden. "Integrating question classification and deep learning for improved answer selection." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3283-3294. 2018.

[5] Kavuk, E. M., & Tosun, A. (2020, June). Predicting Stack Overflow question tags: a multi-class, multi-label classification.In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops* (pp. 489-493).

[6] Manjunath, T. N., Deepa Yogish, S. Mahalakshmi, and H. K. Yogish. "Smart question answering system using vectorization approach and statistical scoring method." *Materials Today: Proceedings* 80 (2023): 3719-3725.

[7] Saedi, Arezoo, Afsaneh Fatemi, and Mohammad Ali Nematbakhsh. "Representation-centric approach for classification of Consumer Health Questions." *Expert Systems with Applications* 229 (2023): 120436.

[8] Chen, Hui, John Coogle, and Kostadin Damevski. "Modeling stack overflow tags and topics as a hierarchy of concepts." *Journal of Systems and Software* 156 (2019): 283-299.

[9] Raida, Maliha Noushin, Zannatun Naim Sristy, Nawshin Ulfat, Sheikh Moonwara Anjum Monisha, Md Jubair Ibna Mostafa, and Md Nazmul Haque. "A study on classifying Stack Overflow questions based on difficulty by utilizing contextual features." *Journal of Systems and Software* 208 (2024): 111884.

[10] Cabrera-Diego, Luis Adrián, Nik Bessis, and Ioannis Korkontzelos. "Classifying emotions in Stack Overflow and JIRA using a multi-label approach." *Knowledge-Based Systems* 195 (2020): 105633.

[11] Hassan, Sk Adnan, Dipto Das, Anindya Iqbal, Amiangshu Bosu, Rifat Shahriyar, and Toufique Ahmed. "Soqde: A supervised learning based question difficulty estimation model for stack overflow." In *2018 25th Asia-Pacific Software Engineering Conference (APSEC)*, pp. 445-454. IEEE, 2018.

[12] WAN, Zhi-yuan, Jia-heng TAO, Jia-kun LIANG, Zhen-gong CAI, Cheng CHANG, Lin QIAO, and Qiao-ni ZHOU. "Large-scale empirical study on machine learning related questions on stack overflow." *Journal of ZheJiang University (Engineering Science)* 53, no. 5 (2019): 819-828.s

[13] Lee, Wai Keat, and Moon Ting Su. "Mining Stack Overflow for API class recommendation using DOC2VEC and LDA." *IET Software* 15, no. 5 (2021): 308-322.

[14] Tóth, László, Balázs Nagy, Dávid Janthó, László Vidács, and Tibor Gyimóthy. "Towards an accurate prediction of the question quality on stack
overflow using a deep-learning-based nlp approach." In *ICSOFT*, pp. 631-639. 2019.

[15] Savenkov, Denis, and Eugene Agichtein. "CRQA: crowd-powered real-time automatic question answering system." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 4, pp. 189-198. 2016.

[16] Beyer, Stefanie, Christian Macho, Massimiliano Di Penta, and Martin Pinzger. "What kind of questions do developers ask on Stack Overflow? A comparison of automated approaches to classify posts into question categories." Empirical Software Engineering 25 (2020): 2258-2301.

[17] Zahedi, Mansooreh, Roshan Namal Rajapakse, and Muhammad Ali Babar. "Mining questions asked about continuous software engineering: A case study of stack overflow." In Proceedings of the 24th International Conference on Evaluation and Assessment in Software Engineering, pp. 41-50. 2020.

[18] Batool, Rabia, Asad Masood Khattak, Jahanzeb Maqbool, and Sungyoung Lee. "Precise tweet classification and sentiment analysis." In 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS), pp. 461-466. IEEE, 2013.

[19] Singla, Saurav, and N. Ramachandra. "Comparative analysis of transformer based pre-trained NLP Models." Int. J. Comput. Sci. Eng 8 (2020): 40-44.

[20] Klein, Tassilo, and Moin Nabi. "Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds." arXiv preprint arXiv:1911.02365 (2019).

[21] Chan, Ying-Hong, and Yao-Chung Fan. "A recurrent BERT-based model for question generation." In Proceedings of the 2nd workshop on machine reading for question answering, pp. 154-162. 2019. G1

[22] Insuasti, Jesus, Felipe Roa, and Carlos Mario Zapata-Jaramillo. "Computers' Interpretations of Knowledge Representation Using Pre-Conceptual Schemas: An Approach Based on the BERT and Llama 2-Chat Models." Big Data and Cognitive Computing 7, no. 4 (2023): 182.

[23] Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).