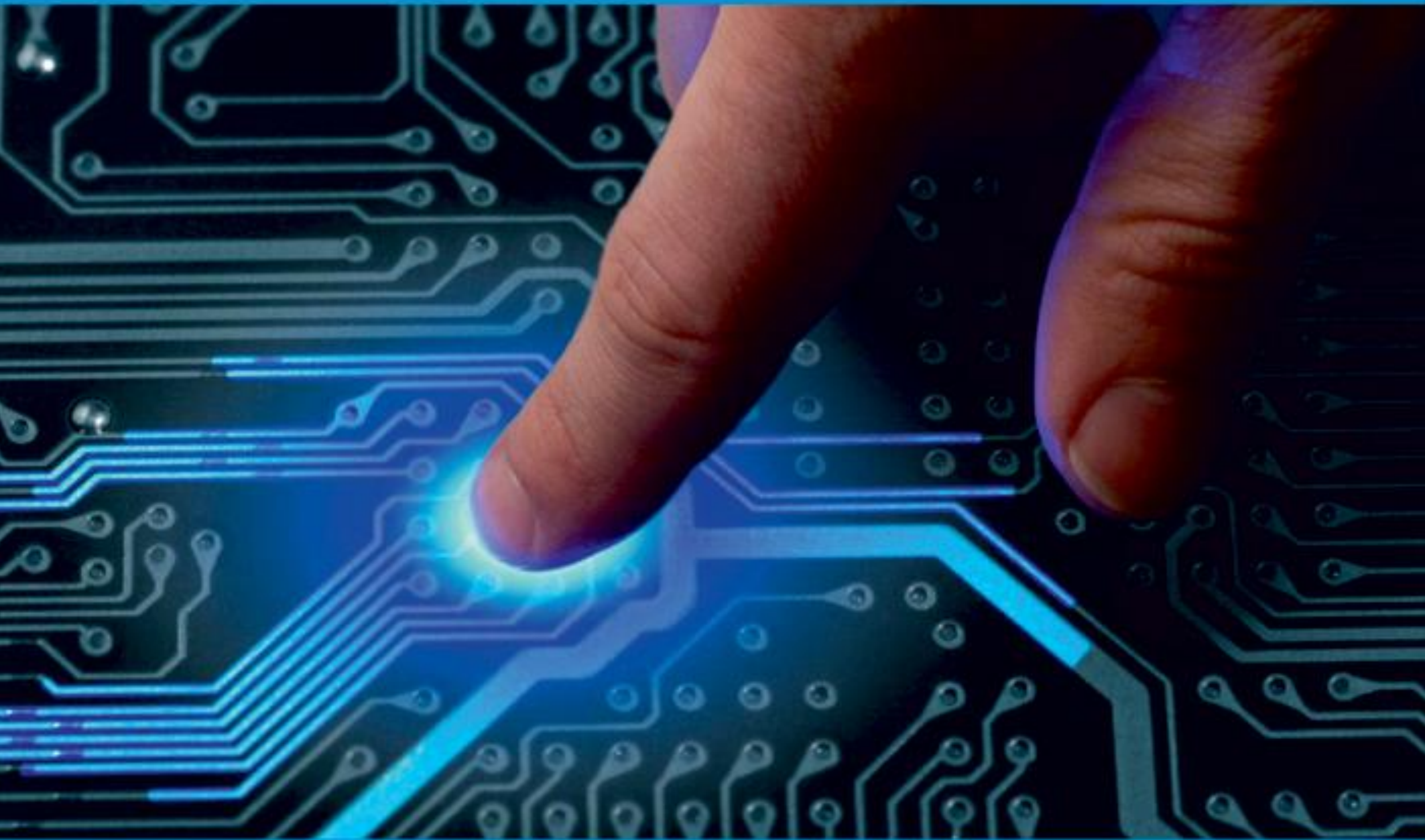
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 5, May 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105001 |

Factoid Question Answering Algorithm

Raymond Zenda[1],Monica Gondo[2], Greenford W Katuruza[3], Paul Chikovo[4], Painos Gweme[5]

Software Engineering Student, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[2]

Software Engineering Lecturer, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[3]

Software Engineering Student, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[4]

Software Engineering Student, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[5]

Software Engineering Student, Dept. of SE., SIST, Harare Institute of Technology, Harare Zimbabwe[5]

ABSTRACT: The amount of information has grown exponentially, making it challenging for search engines to retrieve relevant data. To increase QA accuracy, a neural network factoid question answering algorithm was developed in this study. Using the wikiQA dataset, the method was developed, examined, and tested.The algorithm effectiveness was assessed using statistical assessment measures and our results presented that the algorithm increased accuracy from the RoBERTa-Base+SPP. The proposed study achieved Model Accuracy Performance (MAP) of 93.0% andMean Reciprocal Rank (MRR) of 93.3%. The accuracy of the data from this study is superior to that of Bigram-CNN, LSTM, and RoBERTa+SPP.

KEYWORDS: Natural Language Processing (NLP), Corpus, QuestionAnsweringSystem (QAS), InformationRetrieval (IR), Model Accuracy Performance (MAP),Mean Reciprocal Rank (MRR).

I. INTRODUCTION

In Artificial Intelligence (AI), there is a technique called NLP [1], a method for analysing natural language informationin form of text, sound, visuals and images[2][3].QAS formulates queries posed in natural language and is capable of understanding them[4].Then, using NLP, it extracts the question from the corpus and returns responses[5][6]. In context of IR, QA is the technique of answering natural language queries[7].

Answers are retrieved from a set of natural language documents by QAS[5].It responds to a wide range of queries, including those that are cross-lingual, hypothetical, semantically limited, factual lists, definitions, how, and why[3]. The search collections include regional documents, internal organizational papers, news articles, and the World Wide Web[1]. IR is primarily concerned with the processing, storing, and retrieving numerous forms of the data. In this sector, text retrieval has attracted the most attention[3][1][4][5][7].For many years, retrieving documents with responses to users' natural language questions has been the primary objective of IR. Document retrieval has evolved with the advent of the web and the rapid expansion of search engines[3].Given this avalanche of information, the IR community has focused heavily on accuracy in giving the precise information the user has requested[3][5][7].

A QA system automatically find answers to questions asked in natural language using NLP[4].A QA system may utilise a local or internet database or document collection as the source of the answer. Processing of the Question processing, Retrieval of the Passages, and Processing the final Answer are the three main components of a QAS[8].In the stage of question processing, the user query's requirements and objectives are extracted along with the query's structural data[9][10]. Passage retrieval is process of locating the texts and passages within the text collection that contain the solutions[10]. Answer processing phase is the stage in which the solution is discovered among the possible sentences[10].After the relevant answers that were pulled from the passages have been assessed, most pertinent solution is returned as solution[9].

Major objective of a QAS is to supply the wanted information precisely while requiring minimal human work [10][11]–[13]. Natural language data inquiries are regularly made by users of digital information. Search engines deliver a group of the most documents which are relevant in answer to a question, but user must further investigate

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105001 |

these articles to discover the desired response, which takes a lot of time[9]. In order to get specific information from the corpus that answers the enquiry, IR approaches are used from the standpoint of a QA system [14]. Due to the amount of information that has grown exponentially and readily available on the WWW, search engines have become much more common[1][4]. However, due to information overloading from the amount of information that has grown exponentially, search engines face difficulties in the extraction of pertinent and relevant data.

A factoidQA algorithm is suggested in this study, which crawls through the search engine's results to find the most correct response in the form of a single syllable or word.The system extracts keywords and conducts keyword analysis in order to search for and find the answer[5].

II. Literature Survey

M. Wang, (2006) studied IR techniques for answering factoids. The author examined simillarities and connections of textual implicationalso zeroanaphora resolutionby pertinent work and discovered that dependence structures' full potential for determining QA equivalence haven't been achievedfully. It was learnt that QA systems have relied upon to assess structural matching have mostly remained ad hoc [11].

The factoid QA system that M. Vinodkumar Sadhuram et al. (2020)used the analysis of lexical chain and keywords to developed factoid QAS that used answers factoids from a collection of articles. The author used reasoniung to rate the model perfomance. SQUAD1.0 dataset split at the ratio of 80:20, training and validation respectively was used for validating, training and testing their model. It achieved TF-passage IDF's retrieval accuracy of 69.69% and was averaged at69.93% [5].

In 2020 M. A realistic QA architecture was developed by Y. Day et al. that combined BERT, Q-EAT and AT models for query classification. The authors found that classifying questions and responses can increase the Exact Match (EM) ratio. When the response category and the query are the same, the prediction accuracy improves. [15].

B. Ojokoh et al.,(2021) developed a Bengali language closed domain factoid QAS. The responses retrieved when their model was applied had an accuracy of 66.2% and the response accuracy without the model was 56.8%. Their system managed to retrieve 72% of the documents. The questionclassifiers achieved an accuracy of 90.6% and document classifiers achieved an accuracy of75.3% over five coarse-grained categories.[13].

III. Materials and Methods

A. *Dataset:*

The WikiQA dataset was used in this work for model validation. A new set of questions and sentences that are available to the public, WikiQA dataset, was gathered and annotated for study on question answering[16].In order to accurately reflect the genuine information needs of the general user, questions were derived from Bing query logs. Each query contains a link to a possible solution on a Wikipedia page. The corpus consists of 29,258 sentences and 3,047 questions, with 1,473 sentences designated as answers to the related questions[16], [17], [18]–[25],[26]–[30].

B. *Experimental Setup:*

The algorithm was developed using Pythonlanguage with built-in machine learning dependencies. Hugging Face's Transformers, Matplotlib, Keras, Natural Language Toolkit (NLTK), Scikit, are the primary libraries used in this work.For the experiment we used a laptop running Windows 10 OS with a 500GB hard disk drive and an Intel(R)-Core(TM)i5 CPU @ 1.80GHz and 1.8GHz processing specs.

C. *Preprocessing:*

This phase is a vital component of NLP tasks, which formats text to bemore predictable, understandable and readyfor text classification algorithms[31]. Additionally, pre-processing decreased the feature sparsity of feature representations, making it a tried-and-true technique for enhancing the prediction ability of the classifier algorithms, and shortens the training period by minimizing the required computational resources of a classifier[32][33]. Following is the initial processing of the dataset:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105001 |

Step 1: Cleaning the text:

Eliminate any terminology, tags, and URL prefixes beginning with "https:" and "http://." that conflicted with American National Standards Institute (ANSI) requirements.

Step 2: Tokenization:

The tokenize() technique breaks paragraphs into sentences and sentences into words. For example, “Hey there. How are you” is broken into [‘Hey there’, ‘How are you’] and further broken to [‘Hey’, ‘there’, ‘How’, ‘are’, ‘you’].

Step 3: Stemming:

The phenomena of stemming is the existence of several word forms [34]. Stemming removes a word's suffix using the PorterStemmer() function. For instance, the root word "speak" has the variants "talking," "talks," and "talked."

Step 4: Lower casing:
The sample's word tokens were all converted to lower case.

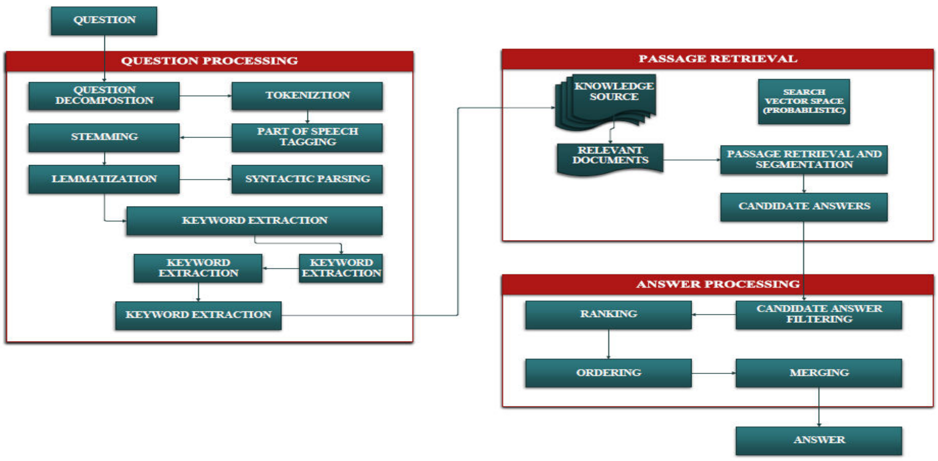
Step 5: Stop word removal:

NLTK was used to eliminate stop words like "the", "is", "are", "it", and "a".

Step 6: Removal of Null values:

All rows with null values were removed using a pandas library to avoid to programming errors due to null values.

D. *Proposed Method:*



Step 1: Question Pre-processing

This is the most crucial tenet to follow when a precise grasp of the knowns and unknowns is the goal. [35]

Question.: “Who is the president of Zimbabwe?”

The grammatical structure needs to be looked at before asking the QAS the aforementioned question. Statistical techniques like Maximum-Entropy-Tagging or Hidden-Markov-Model (HMM) Tagging are used to tokenize the query and assign the proper part of speech tags (POS).

*Who/*WDT, *is/I*N, *the/DT, president/N*N, *of/JJ,Zimbabwe/*NNP, *?/.*

The parts of speech must be allocated in order to perform word-meaning-disambiguation, dependency-parsing, stemming, and Named-Entity-Extraction. The Named-Entity-Recognition (NER), which extracts proper names present in a phrase and categorizing them accordingly to their sort, was utilized in combination with Penn-Treebank-tag-set for English.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105001 |

Zimbabwe.:(GPE.)president.:(ORG.)

Syntactic parsing use parse trees to specifysyntactic structure of a sentence. The dependencies between the tree's nodes were specified using the CLEAR NLP tags. Semantic analysis was utilized to pinpoint aspects like known data, unknown data, and query constraints in order to create a meaningful representation of the query.

In order to retrieve relevant documents, a question built by features collected previouslyis executed on data repositories. A frame-based form with these properties was created and saved.

[person.:[president.: zimbabwe.]]

A comparable question that the system had already successfully responded to was located using classification.

QType..:(PERSON.)(country.)

The below query was created based on the frame-based structure and operators. Boolean, logical, and assignment operators were utilized to show how the constraints, known and unknown characteristics, and both, relate to one another. By connecting the sub-questions of each of the several sub-questions, concatenated questions can be formed. As a result, brackets must yield to logical and boolean operators.

(person.)? (president.)? ==(president.)?(zimbabwe.)?

Step 2: Passage Retrieval

The process of passage retrieval identifies the passages or papers which are most likely to contain the correct answer in the batch of documents[2]. All the relevant documents are retrieved from the corpora when the query is conducted[10]. The employed techniques were the query domain analysis and domain's pages hierarchical clustering. By concentrating on areas that appeared often, the document's dimensionality was significantly decreased.

(person)= (president) =(country)

This approach drastically reduced the search space by focusing on the pertinent document groups and even searching across the hierarchy levels. Relevant excerpts that might help in constructing answers are selected from these texts. We removed passages that don't have plausible solutions. ranking was determined by set of factors, including named entities, search phrases, and their proximity.

{A1:KemboDugishCampbell Mohadi is a Zimbabweanpolitician and former Vice-President of Zimbabwe who served from 28December2017to1March2021.},
{A2:Mnangagwawas sworn in as president of Zimbabwe on 24November2017 at the national sports stadium in Harare before crowd of around 60000},
{A3: Robert. Gabriel. Mugabewas a Zimbabwearevolutionary and politician who served as PrimeMinister of Zimbabwe from 1980 to 1987 and then as President from 1987to2017},
{A4:Emmerson Dambudzo Mnangagwa is a Zimbabweanrevolutionary and politician who has served as President of Zimbabwe since 24November2017.},
{A5:Chiwengathe vice presidentof Zimbabwe isthe 2nd highest political positionobtainable in Zimbabwe}

Vector-Space-Model, was used to represent the relevant queries and documents as vectors of features that indicates terms in relevant documents. This was done for the purposes of IR and response extraction. The search procedure used stemming and lemmatization, which were carried out in the first phase, to provide an exact frequency count for a specific phrase. The procedure next selects the proper response from the group of possible answers produced in the previous stage by using N-gram tiling to produce five potential answers. The initial stage of N gram tiling, known as N gram mining, involved retrieving and weighting each unigram, bigram, and trigram from the filtered passages.

Step 3: Answer Processing

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105001 |

The chosen response is selected from a set of initially discovered potential responses. In this final step of the problem-solving process, each solution on the list of possible solutions must be evaluated and ranked. If you want to know if a response is the best or most correct one that is possible, you need to go back and study the steps that lead to it. Based on the information gathered, an answer can be rejected or accepted and assessed with its accuracy.

1.)A2,A4(0.986,0.92):Emmerson Mnangagwa
2.)A1,(0.5)Kembo Mohadi
3.)A3,(0.4):Robert Mugabe
4.)A5,(0.30):Chiwenga

Using N-gram scores, the candidate replies are sorted and scored at this stage. A candidate solution with highest accuracy is most likely to be the correct answer, and the one with the least accuracy is removed until we end up with one correct answer.

Answer.: "Emmerson Mnangagwa"

Three key processes must be finished before the response can be given in a whole sentence: choosing the topic, organizing the material, and creating the entire phrase. Answers A2 and A4 both returned one fact, although they do so from different sources, whereas response A2 includes two terms that explain a connected reality.The logical next step after choosing the proper material is ordering the information, by combining it into a coherent order usingand Entity-Grid-Representations and Coreference-Based-Coherence algorithms. Mean Reciprocal Rank (MRR), was used to evaluate and rank answers from the first candidate answer set.

IV. RESULTS AND DISCUSSION

We compared the results of our algorithm against results produced byRoBERTa-Base + SPP, RoBERTa-Base Joint MSPP, RE2, Hyper QA, LSTM and Bigram-CNN. The algorithm was tested on WikiQA dataset with training: testing ratio of 80:20. We used two standard functional metrics, MAP and MRR. The model's performance onQA was used toevaluation it.

A. Analysis ofModel Accuracy Performance (MAP)

The comparison results of our algorithm with othersix baseline algorithms are tabulated in Table I.Results shows that our algorithm produced the best MAP of 93%. There is no much difference in accuracy of transformer-based algorithms. Transformer-based algorithms outperformed RE2and Hyper QA algorithms.An accuracy of 65.2% was produced by the least performing algorithm, this is becauseBigram-CNN can't capturelong-term dependencies better than transformers.

Algorithm	MAP	MRR
-ProposedAlgorithm.	0.930.	0.934.
-R0BERTa-Base+SPP.	0.887.	0.899.
-R0BERTa-BaseJointMSPP.	0.885.	0.890.
-RE2.	0.7452.	0.7618.
-HyperQA.	0.712.	0.727.
-L.S.T.M.	0.6552.	0.6747.
-BigramCNN	0.6520.	0.6652.

Table.1:MAP and MRR scores of the QAS algorithms on wikiQA dataset

B. Analysis ofMean Reciprocal Rank (MRR)

Table I: shows that our algorithm and RoBERTa-Base + SPP recorded the highest MRR of 93.4% and 89.9% respectively, whilstleast MRR score of 66.5%was recorded for Bigram-CNN.Despite of the results of Bigram-CNN, it scored an MRR which is greatermodels using classical machine learning. This confirms that the performance of NLP tasks is improved by attention tasks.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105001 |

V. CONCLUSION AND FUTURE WORK

Exponential growth of information on the World Wide Web(WWW) has made QAS become more prevalent. Because of the abundance of information, search engines find it challenging to extract more relevant and contextual data.Thethesisexplored various transformer-based models for answering factoids.We used two state of art metrics, MAP and MRR to evaluate our algorithm's performance.Our algorithm outperformed other algorithmsand we, therefore,conclude that it can improve accuracy of information retrieved by factoid QA. The algorithm searches through the pages and documents that Google search results provide and responds to a query with a single word or sentence. In future work we plan to be able to process a query from images, audios and videos, and also expand the QAS to answer factoid questionsposed in a variety of natural languages other than English.

REFERENCES

[1] B. A. Patanaik, "Open Domain Factoid Question Answering System," no. May, 2018.

[2] R. Qian and X. Hou, "A Knowledge Representation Method for Question Answering Service in Mobile Edge Computing Environment," *Secur. Commun. Networks*, vol. 2022, 2022, doi: 10.1155/2022/1615596.

[3] S. R. M. Aditya, S. R. Rajeswari, D. Reddy, and Varshini, "Factoid question and answering system," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 4, pp. 60–61, 2019.

[4] A. Srm, S. Rajeswari, and D. Reddy, "Factoid Question and Answering System," 2019.

[5] M. Vinodkumar Sadhuram and A. Soni, "Natural Language Processing based New Approach to Design Factoid Question Answering System," *Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2020*, pp. 276–281, 2020, doi: 10.1109/ICIRCA48905.2020.9182972.

[6] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," in *Procedia Computer Science*, 2015, vol. 73, pp. 366–375. doi: 10.1016/j.procs.2015.12.005.

[7] M. Latifi, "Using natural language processing for question answering in closed and open domains," *TDX (Tesis Dr. en Xarxa)*, no. March, pp. 1–165, 2018, [Online]. Available: https://upcommons.upc.edu/handle/2117/118781

[8] A. A. Zulen and A. Purwarianti, "Study and implementation of monolingual approach on indonesian question answering for factoid and non-factoid question," *PACLIC 25 - Proc. 25th Pacific Asia Conf. Lang. Inf. Comput.*, pp. 622–631, 2011.

[9] J. Mozafari, M. Nematbakhsh, and A. Fatemi, "Improved answer selection for factoid questions," *2019 9th Int. Conf. Comput. Knowl. Eng. ICCKE 2019*, no. Iccke, pp. 143–148, 2019, doi: 10.1109/ICCKE48569.2019.8965131.

[10] D. Croce, A. Zelenanska, and R. Basili, *Neural Learning for Question Answering in Italian*, vol. 11298 LNAI. Springer International Publishing, 2018. doi: 10.1007/978-3-030-03840-3_29.

[11] M. Wang, "A survey of answer extraction techniques in factoid question answering," *Comput. Linguist.*, 2006, [Online]. Available: http://www.cs.cmu.edu/~mengqiu/publication/LSII-LitReview.pdf

[12] U. Khanna and D. Mollá, "Transformer-based language models for factoid question answering at BioASQ9b," *CEUR Workshop Proc.*, vol. 2936, pp. 247–257, 2021.

[13] B. Ojokoh *et al.*, "Bengali Question Answering System for Factoid Questions: A statistical approach," *Procedia Comput. Sci.*, vol. 2022, no. 3, pp. 27–28, 2021, doi: 10.1109/ICBSLP47725.2019.201512.

[14] E. Trandafili, E. K. Meçe, K. Kica, and H. Paci, "A novel question answering system for albanian language," *Lect. Notes Data Eng. Commun. Technol.*, vol. 17, pp. 514–524, 2018, doi: 10.1007/978-3-319-75928-9_46.

[15] M. Y. Day and Y. L. Kuo, "A Study of Deep Learning for Factoid Question Answering System," in *Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020*, Aug. 2020, pp. 419–424. doi: 10.1109/IRI49571.2020.00070.

[16] W. Y. C. Meek, "W IKI QA : A Challenge Dataset for Open-Domain Question Answering," no. September 2015, pp. 2013–2018, 2018, [Online]. Available: http://www.aclweb.org/anthology/D15-1237

[17] L. Di Liello, S. Garg, L. Soldaini, and A. Moschitti, "Pre-training Transformer Models with Sentence-Level Objectives for Answer Sentence Selection," no. ii, 2022, [Online]. Available: http://arxiv.org/abs/2205.10455

[18] Z. Wang, H. Mi, and A. Ittycheriah, "Sentence similarity learning by lexical decomposition and composition," *COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap.*, no. challenge 2, pp. 1340–1349, 2016.

[19] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *31st Int. Conf. Mach. Learn. ICML 2014*, vol. 4, pp. 2931–2939, 2014.

[20] Y. Tay, L. A. Tuan, and S. C. Hui, "Hyperbolic representation learning for fast and efficient neural question answering," *WSDM 2018 - Proc. 11th ACM Int. Conf. Web Search Data Min.*, vol. 2018-Febua, pp. 583–591, 2018, doi: 10.1145/3159652.3159664.

[21] T. Munkhdalai and H. Yu, "Neural semantic encoders," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 1, pp. 397–407, 2017, doi: 10.18653/v1/e17-1038.

[22] H. He and J. Lin, "Pairwise word interaction modeling with deep neural networks for semantic similarity measurement," *2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL HLT 2016 - Proc. Conf.*, pp. 937–948, 2016, doi: 10.18653/v1/n16-1108.

[23] L. Di Liello, S. Garg, L. Soldaini, and A. Moschitti, "Paragraph-based Transformer Pre-training for Multi-Sentence Inference," *NAACL 2022 - 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, no. i, pp. 2521–2531, 2022, doi: 10.18653/v1/2022.naacl-main.181.

[24] R. Yang, J. Zhang, X. Gao, F. Ji, and H. Chen, "Simple and effective text matching with richer alignment features," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 4699–4709, 2020, doi: 10.18653/v1/p19-1465.

[25] S. Yoon, F. Dernoncourt, D. S. Kim, T. Bui, and K. Jung, "A compare-aggregate model with latent clustering for answer selection," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 2093–2096, 2019, doi: 10.1145/3357384.3358148.

[26] Z. Ma and M. Collins, "Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 3698–3707, 2018, doi: 10.18653/v1/d18-1405.

[27] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman, "Deep Learning for Answer Sentence Selection," pp. 1–9, 2014, [Online]. Available: http://arxiv.org/abs/1412.1632

[28] C. dos Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive Pooling Networks," no. Cv, 2016, [Online]. Available:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 5, May 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1105001 |

http://arxiv.org/abs/1602.03609

[29] D. Shen *et al.*, “Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms”.

[30] Y. Miao, L. Yu, and P. Blunsom, “Neural variational inference for text processing,” *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 4, no. Mcmc, pp. 2589–2600, 2016.

[31] R. T. Mutanga and S. N. Naicker, “DURBAN UNIVERSITY OF TECHNOLOGY A Comparative Study of Deep Learning Algorithms for Hate Speech Detection on Twitter By A dissertation submitted in fulfilment of the requirement for the Master of Information and Communications Technology degree Faculty of,” 2021.

[32] A. K. Uysal and S. Gunal, “The impact of preprocessing on text classification,” *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014, doi: 10.1016/j.ipm.2013.08.006.

[33] A. I. Kadhim, “An Evaluation of Preprocessing Techniques for Text Classification,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 6, pp. 22–32, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/

[34] T. Tokunaga and M. Iwayama, “Text categorization based on weighted inverse document frequency,” *Tech. Rep. Tokyo Inst. Technol.*, pp. 5–31, 1994.

[35] S. Kadam and A. Gunjal, “A Cognitive approach in Question Answering System,” *Int. J. Eng. Appl. Sci. Technol.*, vol. 2, no. 5, pp. 75–80, 2017.

BIOGRAPHY

Raymond Zendais a Masters student inthe Software Engineering Department, School of Information Science and Technology, Harare Institute of Technology (HIT). He received Bachelor of Technology degree in Information Technology degree in 2019 from HIT, Harare, Zimbabwe. His research interests are Deep Learning, Big Data Analytics, Natural Language Processing etc.

Monica Gondois a Lecturer in the Software Engineering Department, School of Information Science and Technology, Harare Institute of Technology (HIT). Her research interests are Data analytics, Neural networks, Algorithms etc.

Greenford Walter Katuruzais a Software Engineer at Zimbabwe Centre for High performance Computing (ZCHPC) and Space Scientist at Zimbabwe National Geospatial and Space Agency (ZINGSA). He is currently pursuing Master of technology at Harare Institute of Technology, School of Information Science and Technology, Department of Software Engineering, He received Bachelors of Science Honours Geographyin Geospatial Intelligence degree in 2019 from University of Zimbabwe, Harare, Zimbabwe. His research interests are onDeep Learning, AI, Algorithms, Automation, Space technology etc.

Paul Confidence Chikovo is a Masters student in the Software Engineering Department, Harare Institute of Technology. Paul received his Bachelor of Science Honors in Computer Science in 2017 from National University of Science and Technology, Bulawayo, Zimbabwe. His research interests are in NLP, recommender systems, and Deep learning.

Pianos Gweme is a spatial data scientists at the Zimbabwe National Geospatial and Space Agency with vast experience in near real time spatial data collection using UAVs,mobile devices, satellites etc for various applications such as agriculture, mining, disease surveillance, environmental management etc. He is the chief investigator of ZIMSAT-1 the first satellite of Zimbabwe and Chairman of Technical committee of the Zimbabwe center for high performance computing. His research interests are in remote sensing, satellite communication, UAVs software architecture and precision agriculture using UAVs.