GSJ: Volume 10, Issue 12, December 2022
ISSN 2320-9186

Global Scientific JOURNALS

GSJ: Volume 10, Issue 12, December 2022, Online: ISSN 2320-9186
www.globalscientificjournal.com

AN IMPROVED INFORMATION RETRIEVAL MODEL FOR QUERR-BASED MULTIDOCUMENT SUMMARIZATION

ADEWOLE KAZEEM BABATUNDE

ABSTRACT

Information Retrieval (IR) is a system concerned with the retrieval of documents from (stream of) document databases, based on user information. IR carry out its operations by first preparing the collection of documents for retrieval through an indexing process after which user (information) needs are captured by phrases which are themselves indexed and used to rank the documents retrieved. It is functional also in the area of huge scientific literature, in supporting complex software engineering processes such as code maintenance and bug management, and most importantly in automatic document summarization.The growth in electronically available documents makes it difficult to obtain the necessary information related to the needs of a user. Text summarization systems extract brief information from a given document. Summarization is a process where the most relevant features of a text are extracted and compiled into a short abstract of the original document. This work therefore develop a query-based multi-document summarization system leveraging on combination of IR techniques, clustering and graph-based model. This reduces redundancies, summarizes multiple documents with related topics based on semantic (meaning of words) centered on a query input. This study is to provide accurate and useful summarization of the contents of multiple text documents using a user input. With automatic text summarization, this prob-

lem can be mitigated, thereby acquiring more information leading to more effective decisions in less time. Besides, summarization can also benefit other NLP (Natural Language Processing) tasks. So, summarization focuses on getting the main meaning of a document.

1. INTRODUCTION

With the dramatic growth of the Internet, people are overwhelmed by the tremendous amount of online information and documents. This expanding availability of documents has demanded exhaustive research in the area of automatic text summarization. According to an *Approach Design and Analysis of Resource Management Support Software for Multihoming in-Vehicle of IPv6 Network.* [1]. *An Intelligent Spam-Scammer Filter Mechanism Using Bayesian Techniques* in [2] is required.A *summary* is defined as "a text that is produced from one or more texts that conveys important information in the original text, and significantly less than half of the original text. *Automatic text summarization* is the task of producing a concise and fluent summary while preserving key information content and overall meaning. In recent years, numerous approaches have been developed for automatic text summarization and applied widely in various domains. *Character Proximity For RFID Smart Certificate System: A Revolutionary Security Measure to Curb Forgery Menace* in [3].For example, search engines generate snippets as the previews of the documents. *Big Data: A Computing Model for Knowledge Extraction on Insurgency Management* in [4] can provide the dataset, In general, there are two different approaches for automatic summarization: *extraction* and *abstraction*. *Extractive summarization* methods work by identifying important sections of the text and generating them verbatim; thus, they depend only on extraction of sentences from the original text. In contrast, *abstractive summarization* methods aim at producing important material in a new way. In other words, they interpret and examine the text using advanced natural language techniques in order to generate a new shorter text that conveys the most critical information from the original text. Even though summaries created by humans are usually not extractive, most of the summarization research today have focused on extractive summarization. Purely extractive summaries often times give better results compared to automatic abstractive summaries [5]. It exploits the property of sentence ranking methods in which they consider neural query ranking and query-focused ranking. In [6] developed a query-based summarization that uses document ranking, time-sensitive queries and ranks recency sensitive queries as the features for text summarization. ***Energy Efficient Hierarchical Cluster Head Election Using Exponential Decay Function Prediction in [7] will predict the decay function,*** the review in [8] designed a system for automatic keyword extraction for text summarization in single document e-Newspaper article.

2. RELATED WORKS

This type of learning techniques used labeled dataset for training.The information in [9] designed a system for automatic keyword extraction for text summarization using hidden Markov model. The review in [10] used a set of labeled dataset to train the system for the classification of temporal relations between events. *Immune Inspired Concepts Using Neural Network for Intrusion Detection in Cybersecurity in [11].*Bag-of-words model is built at sentence level, with the usual weighted term-frequency and inverse sentence frequency paradigm (Rene Arnulfo et al 2009)[12], where sentence-frequency is the number of sentences in the document that contain that term. These sentence vectors are then scored by similarity to the query and the highest scoring sentences are picked to be part of the summary. This is a direct adaptation of Information Retrieval paradigm to summarization.The information in [***13] shows Zero Day Attack Prediction with Parameter Setting Using Bi Direction Recurrent Neural Network in Cyber Security.*** Documents are represented using term frequency inverse document frequency (TF-IDF) of scores of words [14]. Term frequency used in this context is the average number of occurrences (per document) over the cluster. IDF value is computed based on the entire corpus. ***The review in [15] on Route Optimization in MIPv6 Experimental Test bed for Network Mobility: Trade off Analysis and Evaluation.*** Graph theoretic representation [16] of passages provides a method of identification of these themes. After the common preprocessing steps, namely, stop word removal and stemming, sentences in the documents are represented as nodes in an undirected graph. ***Dynamic Flow Reduction Scheme Using Two Tags Multi-protocol Label Switching (MPLS) in Software Define Network*** in [17] ***and [18] in Prediction of Breast Cancer Images Classification Using Bidirectional Long Short Term Memory and Two-Dimensional Convolutional Neural Network***. Training is performed using a modified version of the back propagation algorithm (Orr &Müller 1999) which is based on the gradient descend method described in [19]. Other summarization systems that also employ neural networks in their algorithms. An extraction-based summarization technique using k-means clustering algorithm which is an unsupervised learning technique. The score for each sentences computed and centroid based clustering is applied on the sentences and extracting important sentences as part of summary**.** *Investigating Data Mining Trend in Cybercrime Among Youths*. Pervasive Computing and Social Networking in [20] and [21] *in Development of DDoS Attack Detection Approach in Software Defined Network Using Support Vector Machine Classifier*. The author performed evaluations against several baselines that include the combination of some of the mentioned features. The overall results he obtained indicate the prevalence of the used decision-tree method. Other variations of the tree-based method were also used. Neural networks are non-linear statistical data models used to model complex relationships between inputs and outputs and detect patterns within text. The structure of 30 the networks usually changes based on the information processed during the training stage. In DUC 2007, a summarizer called NetSum using neural networks-based algorithm has demonstrated a performance significantly exceeding those of the provided baselines. The algorithm used a pair-based sentences ranker called RankNetfor ranking all sentences in the source documents based on their importance.

3. DEVELOPED MODEL

Reading and preprocessing document from plain text files which includes tokenization, stop words removal, case change and stemming. Document Clustering of input documents to group similar documents in clusters.Topic Modelling. IR technique using VSM based on TF-IDF is used for topic modelling. Relevant Documents retrieval against input topics and subtopics. The similarity is to be measured between input topic and topics modellig output to identify most relevant cluster using cosine similarities.Summarization using ‘TextRank’ approach to model text as graph networks and retrieve high importance sentences as summaries. Figure 1 below is a flowchart that depicts the steps taken to generate summary. The stages invoved in the model are as follows:

Stage 1: Statistical natural language processing tools are used in the preprocessing stage to filter out stop list words and generate stem words, by avoiding the inflectional forms of terms. The resulting meaningful stems are very useful during the normalization of terms in the term distribution model.

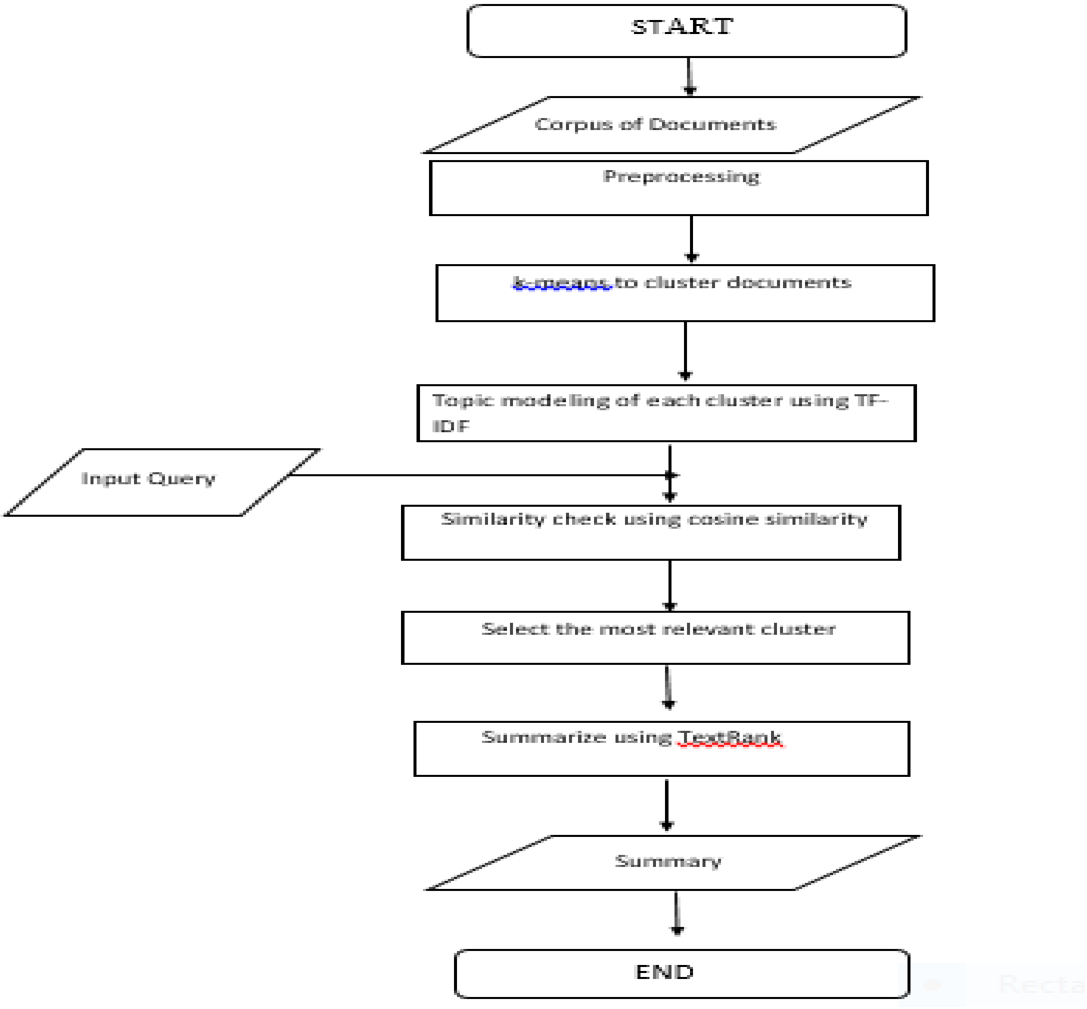

START
Corpus of Documents
Preprocessing
k-means to cluster documents
Topic modeling of each cluster using TF-IDF
Input Query
Similarity check using cosine similarity
Select the most relevant cluster
Summarize using TextRank
Summary
END

Figure1: A flowchart of the Developed model

GSJ: Volume 10, Issue 12, December 2022
ISSN 2320-9186

Stage 2: Stop words are the words occurring very frequently and not conveying independent meaning in a document. For instance, “the”, “would”, “can”, “do” are typical stop words. Prepositions, pronouns, articles, connectives etc. are also considered as stop words. Since they carry very little information about the contents of a document, it is usually a good idea to remove them from the document collections. A list of these words was made and was removed from the documents.

Stage 3: Stemmers are used in Information Retrieval to reduce the size of index files. Since a single stem typically corresponds to many full terms, compression factors of over 50% can be achieved by storing stems instead of terms, especially in the case of affix-removal stemmers. Words such as “wait”,”waiting”,”waits”,”waited” invariably mean the same thing and as such are all converted to a single base word.

Stage 4: K-means clustering algorithm is employed in our approach due to its simplicity, K data elements are selected as initial centers; then distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate cluster. Kmeans is used to form clusters so that related documents are grouped together. In other rwords Kmeans algorithm work to uptimize the minimum, square of Euclidean disatance between member of the same clusters using the equation below:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2$$

The IDF (inverse document frequency) of a word is the measure of how significant that term is in the whole corpus, Similarity is used as a similarity measure after the text is converted to a vectors using VSM model while the Text summarization is the problem of creating a short, accurate, and fluent summary of a longer text document.

Stage 5: TextRank is a graph based algorithm that uses the intuition behind the PageRank algorithm to rank sentences. It is an unsupervised method for computing the extractive summary of a text which mean no training is necessary for the algorithm to work. The text units in the top-ranking list are considered to be keywords of the text. The flowchart of the algorithm is shown below.

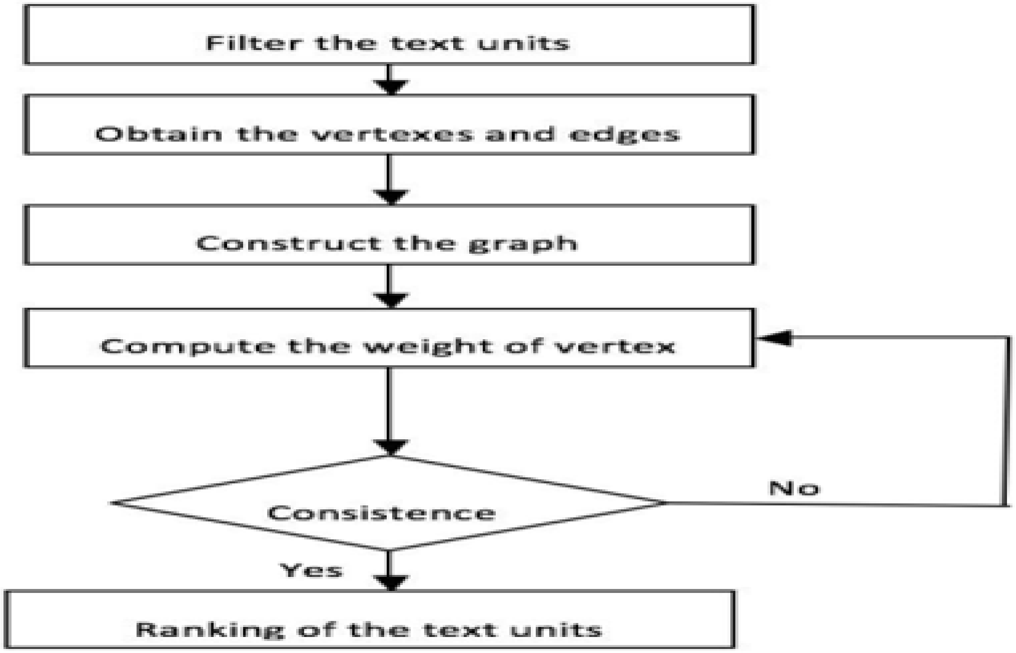
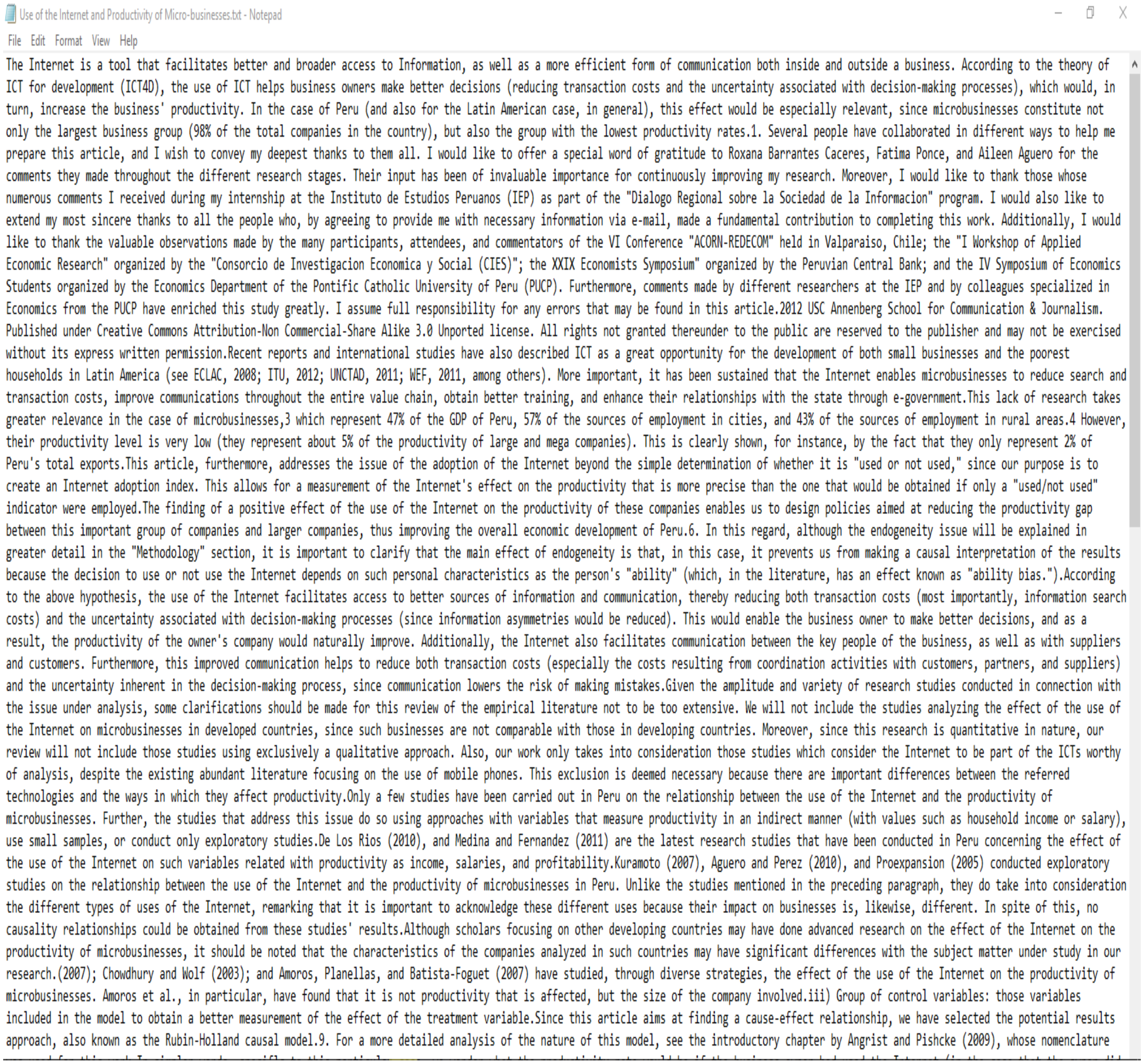

Filter the text units
Obtain the vertexes and edges
Construct the graph
Compute the weight of vertex
Consistence
No
Yes
Ranking of the text units

Figure 2: A flowchart of the TextRank algorithm

The graph generated is symmetrical. Therefore there is a need for a strong similarity measure on which the algorithm's performance will heavily depend. The algorithm is language agnostic. It doesn't require any training, it's a totally unsupervised method.

4. RESULTS AND DISCUSSION

In section, we will be discussing results obtained from the study. The python API Sublime text is used as the environment for the implementation. Sublime text is a source code editor developed by Microsoft for Windows, Linux and macOS. It includes support for debugging, embedded Git control, syntax highlighting, intelligent code completion, snippets, and code refactoring. It is also customizable, so users can change the editor's theme, keyboard shortcuts, and preferences. It is free and open-source, although the official download is under a proprietary license. Thirty different documents downloaded from Onlinessays.com were used to test the model. These documents are centralized around topic finance and economy. The documents were preprocessed before added to the document corpus for summarization. An example of the raw text used in the experiment is given in figure 3.


Use of the Internet and Productivity of Micro-businesses.txt - Notepad
File
Edit
Format
View
Help

The Internet is a tool that facilitates better and broader access to Information, as well as a more efficient form of communication both inside and outside a business. According to the theory of ICT for development (ICT4D), the use of ICT helps business owners make better decisions (reducing transaction costs and the uncertainty associated with decision-making processes), which would, in turn, increase the business' productivity. In the case of Peru (and also for the Latin American case, in general), this effect would be especially relevant, since microbusinesses constitute not only the largest business group (98% of the total companies in the country), but also the group with the lowest productivity rates.1. Several people have collaborated in different ways to help me prepare this article, and I wish to convey my deepest thanks to them all. I would like to offer a special word of gratitude to Roxana Barrantes Caceres, Fatima Ponce, and Aileen Aguero for the comments they made throughout the different research stages. Their input has been of invaluable importance for continuously improving my research. Moreover, I would like to thank those whose numerous comments I received during my internship at the Instituto de Estudios Peruanos (IEP) as part of the "Dialogo Regional sobre la Sociedad de la Informacion" program. I would also like to extend my most sincere thanks to all the people who, by agreeing to provide me with necessary information via e-mail, made a fundamental contribution to completing this work. Additionally, I would like to thank the valuable observations made by the many participants, attendees, and commentators of the VI Conference "ACORN-REDECOM" held in Valparaiso, Chile; the "I Workshop of Applied Economic Research" organized by the "Consorcio de Investigacion Economica y Social (CIES)"; the XXIX Economists Symposium" organized by the Peruvian Central Bank; and the IV Symposium of Economics Students organized by the Economics Department of the Pontific Catholic University of Peru (PUCP). Furthermore, comments made by different researchers at the IEP and by colleagues specialized in Economics from the PUCP have enriched this study greatly. I assume full responsibility for any errors that may be found in this article.2012 USC Annenberg School for Communication & Journalism. Published under Creative Commons Attribution-Non Commercial-Share Alike 3.0 Unported license. All rights not granted thereunder to the public are reserved to the publisher and may not be exercised without its express written permission.Recent reports and international studies have also described ICT as a great opportunity for the development of both small businesses and the poorest households in Latin America (see ECLAC, 2008; ITU, 2012; UNCTAD, 2011; WEF, 2011, among others). More important, it has been sustained that the Internet enables microbusinesses to reduce search and transaction costs, improve communications throughout the entire value chain, obtain better training, and enhance their relationships with the state through e-government.This lack of research takes greater relevance in the case of microbusinesses,3 which represent 47% of the GDP of Peru, 57% of the sources of employment in cities, and 43% of the sources of employment in rural areas.4 However, their productivity level is very low (they represent about 5% of the productivity of large and mega companies). This is clearly shown, for instance, by the fact that they only represent 2% of Peru's total exports.This article, furthermore, addresses the issue of the adoption of the Internet beyond the simple determination of whether it is "used or not used," since our purpose is to create an Internet adoption index. This allows for a measurement of the Internet's effect on the productivity that is more precise than the one that would be obtained if only a "used/not used" indicator were employed.The finding of a positive effect of the use of the Internet on the productivity of these companies enables us to design policies aimed at reducing the productivity gap between this important group of companies and larger companies, thus improving the overall economic development of Peru.6. In this regard, although the endogeneity issue will be explained in greater detail in the "Methodology" section, it is important to clarify that the main effect of endogeneity is that, in this case, it prevents us from making a causal interpretation of the results because the decision to use or not use the Internet depends on such personal characteristics as the person's "ability" (which, in the literature, has an effect known as "ability bias.").According to the above hypothesis, the use of the Internet facilitates access to better sources of information and communication, thereby reducing both transaction costs (most importantly, information search costs) and the uncertainty associated with decision-making processes (since information asymmetries would be reduced). This would enable the business owner to make better decisions, and as a result, the productivity of the owner's company would naturally improve. Additionally, the Internet also facilitates communication between the key people of the business, as well as with suppliers and customers. Furthermore, this improved communication helps to reduce both transaction costs (especially the costs resulting from coordination activities with customers, partners, and suppliers) and the uncertainty inherent in the decision-making process, since communication lowers the risk of making mistakes.Given the amplitude and variety of research studies conducted in connection with the issue under analysis, some clarifications should be made for this review of the empirical literature not to be too extensive. We will not include the studies analyzing the effect of the use of the Internet on microbusinesses in developed countries, since such businesses are not comparable with those in developing countries. Moreover, since this research is quantitative in nature, our review will not include those studies using exclusively a qualitative approach. Also, our work only takes into consideration those studies which consider the Internet to be part of the ICTs worthy of analysis, despite the existing abundant literature focusing on the use of mobile phones. This exclusion is deemed necessary because there are important differences between the referred technologies and the ways in which they affect productivity.Only a few studies have been carried out in Peru on the relationship between the use of the Internet and the productivity of microbusinesses. Further, the studies that address this issue do so using approaches with variables that measure productivity in an indirect manner (with values such as household income or salary), use small samples, or conduct only exploratory studies.De Los Rios (2010), and Medina and Fernandez (2011) are the latest research studies that have been conducted in Peru concerning the effect of the use of the Internet on such variables related with productivity as income, salaries, and profitability.Kuramoto (2007), Aguero and Perez (2010), and Proexpansion (2005) conducted exploratory studies on the relationship between the use of the Internet and the productivity of microbusinesses in Peru. Unlike the studies mentioned in the preceding paragraph, they do take into consideration the different types of uses of the Internet, remarking that it is important to acknowledge these different uses because their impact on businesses is, likewise, different. In spite of this, no causality relationships could be obtained from these studies' results.Although scholars focusing on other developing countries may have done advanced research on the effect of the Internet on the productivity of microbusinesses, it should be noted that the characteristics of the companies analyzed in such countries may have significant differences with the subject matter under study in our research.(2007); Chowdhury and Wolf (2003); and Amoros, Planellas, and Batista-Foguet (2007) have studied, through diverse strategies, the effect of the use of the Internet on the productivity of microbusinesses. Amoros et al., in particular, have found that it is not productivity that is affected, but the size of the company involved.iii) Group of control variables: those variables included in the model to obtain a better measurement of the effect of the treatment variable.Since this article aims at finding a cause-effect relationship, we have selected the potential results approach, also known as the Rubin-Holland causal model.9. For a more detailed analysis of the nature of this model, see the introductory chapter by Angrist and Pishcke (2009), whose nomenclature

Figure3: A sample of a document from the corpus

4.1: Results and Discussion

The model is executed using 20 different queries that cut across all topics discovered in the corpus. Summaries are evaluated based on evaluation metrics precision, recall and F-measure. In other to evaluate the scalability of the proposed approach the queries are grouped into sections as follows: Q1 to Q6 are simple queries containing two words phrase, Q7 to Q15 are moderate queries composed of simple sentence like input and Q16 to Q20 are complex queries. Table 1 shows the results of the metrics on the 20 queries.

GSJ: Volume 10, Issue 12, December 2022
ISSN 2320-9186

Table 1: the results of the metrics on the 20 queries

Queries	Precision	Recall	F-Measure
Q1	0.79	0.87	0.83
Q2	0.75	0.85	0.80
Q3	0.78	0.89	0.83
Q4	0.81	0.81	0.81
Q5	0.70	0.80	0.75
Q6	0.76	0.88	0.82
Q7	0.81	0.87	0.84
Q8	0.85	0.81	0.83
Q9	0.91	0.89	0.90
Q10	0.81	0.84	0.82
Q11	0.79	0.87	0.83
Q12	0.84	0.88	0.86
Q13	0.82	0.90	0.86
Q14	0.80	0.91	0.85
Q15	0.85	0.88	0.86
Q16	0.65	0.78	0.71
Q17	0.69	0.81	0.75
Q18	0.60	0.76	0.67
Q19	0.68	0.84	0.75
Q20	0.67	0.68	0.67

From the results it can be observed that generally the proposed approach generates good results with considerably high F-measure values in all considered queries except in the complex queries where values of precisions are very low. The model gave the peak F-measure value 0.90 in Q9 under moderate queries section. This is so due to the high precision and recall values. Also it can be observed that the precision value of the model increases from simple queries to moderate queries and drop sharply from moderate to complex queries. This moderate scalability results might be due to the simplicity of the approach. For the recall, the proposed model produced a relatively constant value not less than 0.75 in all queries. This underscores the effectiveness of this approach. The summaries created are not more than sixteen (16) sentences with an Accuracy of 87.4%.

5. CONCLUSION

Corpus of document was created and grouped into cluster using K-means which leverages topic model extracted from each document using TF-IDF. Vector space model and Cosine similarity are employed to determine similarities between user query and clusters representatives in order to detmine the most relevant cluster to summarize. An algorithms are implemented in Python and tested on thirty documents related to finance and economy. Performance of the model is evaluated using precision recall and F-measure. Results from experiments show that the proposed summarization can produce good summaries with F-measure as high as 9.0 and accuracy of 87%. It is obvious that the approach can be applies to simple and moderate queries effectively. This propose model however could not effectively handle complex queries due to simplicity of the approach. This was accomplished by using TextRank.

References

[1] S.D. Adeniji. S. Khatun, M.A. Borhanudin. R.S.A.Raja. "*An Approach Design and Analysis of Resource Management Support Software for Multihoming in-Vehicle of IPv6 Network*". Journal of International Review on Computers and Software RECOS vol. 3. No. 193-197. 2008.

[2] O.D. Adeniji, O. Adigun, O. O. Adeyemo. "*An Intelligent Spam-Scammer Filter Mechanism Using Bayesian Techniques*". International Journal of Computer Science and Information Security IJCSIS Vol. 10. No. 3. pp 126-13 . 2012.

[3] C. Eze, O.D. Adeniji. "*Character Proximity For RFID Smart Certificate System: A Revolutionary Security Measure to Curb Forgery Menace*". International Journal of Scientific and Technology Research IJSTR, October,08, Vol 3 No 66-70. 2014.

[4] O.D. Adeniji. "*Big Data: A Computing Model for Knowledge Extraction on Insurgency Management*". International Conference on Information and Communication Technology and Its Application. Pp52-55. 2016

[5] D. Hingu "Automatic text summarization of Wikipedia articles," in Communication, Information & Computing Technology (ICCICT), International Conference on. IEEE, , pp. 1–4. 2015

[6] D. Wang. "Multi-document summarization using sentence-based topic model". In Proceedings of the ACL-IJCNLP , 2009 .

[7] A.O. Ojoawo. O.D. Adeniji. *Energy Efficient Hierarchical Cluster Head Election Using Exponential Decay Function Prediction.* International Journal of Wireless & Mobile Networks (IJWMN), Vol. 10, No. 5. pp 17-31. 2018

[8] L. Zhou. “Summarizing Answers for Complicated Questions“. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006), Genoa, Italy

[9] Jiayin . “Approaches to Event-Focused Summarization Based on Named Entities and Query Words“. In Proceedings of Document Understanding Conferences, 2003.

[10] Baxendale, P. B. (1958). Man-made index for technicalliterature - An experiment. IBM Journal of Research and Development, 2(4):354-361

[11] O.D. Adeniji, J.J Ukam. *Immune Inspired Concepts Using Neural Network for Intrusion Detection in Cybersecurity.* Proceedings of the 20th iSTEAMS Multidisciplinary Trans-Atlantic Going Global Conference. pp 119-126, 2019.

[12] Mihalcea R, Tarau P .TextRank: Bringing Order into Texts [J]. UNT, 90:404-411.2004.

[13] . O.D. Adeniji, O.O. Olatunji. *Zero Day Attack Prediction with Parameter Setting Using Bi Direction Recurrent Neural Network in Cyber Security.* International Journal of Computer Science and Information Security IJCSIS, March, 03, Vol. 18. No. 3. pp 111-118. 2020.

[14] E. J. Santo. "Automatic Evaluation of Summaries Using Document Graphs," presented at Text Summarization Branches Out: Proceedings ofthe ACL-04 Workshop, Barcelona, Spain, pp.66-73, 2004.

[15] O.D. Adeniji, A. Osofisan. "*Route Optimization in MIPv6 Experimental Test bed for Network Mobility: Trade off Analysis and Evaluation*". *International Journal of Computer Science and Information Security* IJCSIS,Vol. 18. No. 5. pp 19-28. 2020.

[16] D.R.Radev. "Centroid-based summarization of multiple documents: sentence extraction, utility based evaluation, and user studies". Proceedings of the 2000 NAACL ANLP Workshop on Automatic summarization - Volume 4. Seattle, Washington, Association for Computational Linguistics: 21-30.

[17] O.D. Adeniji *.* "*Dynamic Flow Reduction Scheme Using Two Tags Multi-protocol Label Switching (MPLS) in Software Define Network*". International Journal of Emerging Trends in Engineering Research. March, 03, Volume 10. No.3, 2022.

[18] I.R.Idowu, O.D. Adeniji, O.D. S. Elelu, S., & T.O .Adefisayo. "*Prediction of Breast Cancer Images Classification Using Bidirectional Long Short Term Memory and Two-Dimensional Convolutional Neural Network*". Transaction Networks and Communications. August, 25, Vol.9, No.4. UK. (2021).

[19] akeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and ShinIshii. Distributional Smoothing with Virtual Adversarial Training.arXiv:1507.00677(2016). (PDF) Adequacy of the Gradient-Descent Method for Classifier Evasion Attacks. Available from: https://www.researchgate.net/publication/315807236_Adequacy_of_the_Gradient-escent_Method_for_Classifier_Evasion_Attacks [accessed Dec 29 2022].

[20] A.O Adesina , S.A. Ajagbe , O.S Afolabi , O.D. Adeniji , and O.I. Ajimobi. “Investigating *Data Mining Trend in Cybercrime among Youths*". Pervasive Computing and Social Networking. Lecture Notes in Networks and Systems, vol 475.pp 725-741, Springer, Singapore.2022.

[21] O.D. Adeniji , D.B Adekeye, S.A Ajagbe, A.O. Adesina, Y.J Oguns, M.A. Oladipupo. "*Development of DDoS Attack Detection Approach in Software Defined Network Using Support Vector Machine Classifier*". In: Ranganathan, G., Bestak, R., Fernando, X. (eds) Pervasive Computing and Social Networking. Lecture Notes in Networks and Systems, vol 475.pp319-331, Springer, Singapore.2022.