Extractive Summarization on Twi Language using Text-Rank and Fuzzy Clustering Algorithm

1st Sayak Dhole
School of Computer Engineering
KIIT Deemed to be University
India
2028155@kiit.ac.in

2nd Santwana Sagnika
School of Computer Engineering
KIIT Deemed to be University
India
santwana.sagnikafcs@kiit.ac.in

3rd Ampofo Dominic
School of Computer Applications
KIIT Deemed to be University
India
asadom2019@gmail.com

4th Bhabani Shankar Prasad Mishra
School of Computer Engineering
KIIT Deemed to be University
India
bsmishrafcs@kiit.ac.in

5th Aloka Fernando
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
alokaf@cse.mrt.ac.lk

6th Satya Ranjan Dash
School of Applications
KIIT Deemed to be University
India
sdashfca@kiit.ac.in

2023 1st International Conference on Circuits, Power and Intelligent Systems (CCPIS) | 979-8-3503-1576-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCPIS59145.2023.10291396

***Abstract*—Twi is a dialect of the Akan language spoken in southern and central Ghana by several million people. About 80% of the Ghanaian population speaks Twi as a first or second language. Summarization is a basic natural language processing task, that is a step to further more advance tasks as well as an aid in reviewing information. This paper tackles the problem of summarizing Twi texts. Three extractive summarization techniques - TF-IDF, Text-Rank and Fuzzy C-means Clustering are implemented. We find that Text Rank performs best, while TF-IDF despite being fairly simple performs comparably well to Text Rank. Our work suggests that embedding based approaches are potent and further work should employ their use.**

***Index Terms*—Twi Language, Text-Rank, Fuzzy Clustering, TF-IDF, Summaization.**

[1]

I. Introduction

Twi is a dialect of the Akan language, which is spoken by the Akan people in Ghana. It is a tonal language, which means that the tone or pitch of a word can change its meaning. Twi has two main dialects: Asante Twi and Akuapem Twi. Asante Twi is the more widely spoken of the two dialects and is considered the prestige dialect. Twi is one of the most widely spoken languages in Ghana, with around 7 million speakers. The Twi language has a relatively simple grammar compared to English, with no gender distinctions or verb conjugation. Instead, Twi relies on word order and particles to convey meaning. For example, in Twi, the particle ”na” is used to indicate that a noun is the subject of a sentence. Twi also has a rich and varied vocabulary, which includes many proverbs and idiomatic expressions. Proverbs are essential part of Akan culture and are often used in everyday conversation to convey wisdom and advice.

The history of the Twi language can be traced back to the Akan people, who are believed to have migrated to what is now Ghana from the area that is now known as the Ivory Coast in the 11^{th}century. The Akan people established several powerful states in the region, including the Ashanti Empire, which was one of the most powerful states in West Africa in the 18^{th} and 19^{th} centuries. During the colonial period, English became the dominant language in Ghana, and many Ghanaians began to learn English as a second language. However, Twi and other indigenous languages remained important in everyday life, and they continue to be spoken today.

In recent years, there has been a renewed interest in the Twi language and Akan culture in Ghana. Efforts have been made to promote the language and preserve it for future generations. Twi is now taught in many schools in Ghana, and there are also several online resources available for learning the language. The Akan people are proud of their language and culture, and they see it as an important part of their identity. Twi has about 17–18 million speakers in total, including second-language speakers; about 80% of the Ghanaian population speaks Twi as a first or second language. Twi is a common name for mutually intelligible former literary dialects of the Akan language, Bono, Asante, and Akuapem. It is also a low resource language, with limited datasets. As a result, research on performing NLP tasks on datasets of this language is limited. In this paper, we attempt to bridge the gap by demonstrating three methods of Extractive Summarization on Twi texts. Extractive Summarization is a method to summarize a text by selecting important sentences. It is a crucial task in the age of information, since it is difficult to assess the relevance of a text without reading.
Extractive summarization is a popular technique because it is relatively simple and produces summaries that are faithful to the original text. It involves selecting important sentences

[1]979-8-3503-1576-9/23/$31.00 ©2023 IEEE

or phrases from the original text and combining them to form a summary. Extractive summarization methods often use statistical techniques such as sentence ranking or clustering to identify the most important sentences in the text. This method is based on the assumption that the most important information in a text is often contained in its key sentences [3]. Extractive summarization can be viewed as a text classification problem, where each sentence in the input text is assigned a score based on its importance, and the most highly ranked sentences are selected to form the summary. Extractive summarization techniques typically rely on statistical methods such as frequency-based methods, graph-based methods, and machine learning approaches.

In this paper, we implement three methods to carry out Extractive Summarization. They involve assessing individual words of a sentence in order to measure the importance of the sentence. The first method is TF-IDF, where a score based on term frequency is calculated to determine the importance of a word. Second Method is TextRank which treats each sentence within a text as nodes in a graph. Weight of each edge is calculated using cosine similarity of the sentence embeddings.

II. Literature Review

Considering the scarcity of work done on Twi language texts, especially any demonstration of Extractive Summarization, we explored what methods have given good results on other African languages. TF-IDF has been shown to give good results for extractive summarization of Kiswahili texts [2]. Human evaluation was employed in the aforementioned study, which we use as well. TF-IDF has also shown promise in other non-English languages [9].

Text Rank is a very wide spread approach to text summarization and is language agnostic given one generates embeddings of the text. It has been used with good results for non-English languages which has been the motivation behind selecting this method [4]. // Since there is very little language-knowledge and a lack of labelled data, unsupervised Machine Learning methods are necessary as they require no labelling of data. Fuzzy C-means is one such method that employs fuzzy logic to apply clustering to the sentences within a text. It selects the sentences which "belong" the most to the cluster. It has not been applied to non-English languages in the past. However, it's shown promise on the standard extractive summarization dataset : CNN/Daily Mail summarization dataset [8], to obtain a ROUGE F measure of 0.53. The constraints of limited data can be handled with this method [6].

Extractive summarization plays a crucial role in distilling information from textual sources. This task has gained increasing attention, leading to the development of summarization techniques tailored for specific languages. In the context of the Twi language, researchers have focused on developing extractive summarization models to address the need for automated summarization in this language.

Another approach is Fuzzy C-Means (FCM), a clustering algorithm that has been utilized in Twi extractive summarization. FCM clusters similar sentences together based on linguistic features, allowing the identification of key themes or topics within the text. By grouping sentences into representative clusters, FCM assists in selecting important sentences for the summarization process.

Additionally, TextRank, a graph-based ranking algorithm, has been adapted for Twi extractive summarization. TextRank constructs a sentence graph, where each sentence represents a node, and the edges represent the similarity or relatedness between sentences. By applying iterative ranking algorithms, such as random walks, TextRank identifies the most important sentences in the graph, thus generating the summary [10].

To assess the quality of Twi extractive summarization, evaluation metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) have been utilized. These metrics measure the overlap between the generated summaries and reference summaries, enabling researchers to evaluate the effectiveness of their models.

Various approaches and techniques to tackle Twi extractive summarization. These methods include statistical models, graph-based algorithms, machine learning approaches, and linguistic feature-based methods. For instance, the LexRank algorithm has been utilized to generate summaries in the Twi language. Comparative analyses of existing Twi extractive summarization systems, have explored the performance, limitations, and practical applications of different approaches. Additionally, case studies have demonstrated the utility of Twi summarization in domains such as news articles, social media, and educational content.

III. Methodology

Data was collected from people who speak Twi, 22 paragraphs were submitted which were pre cleaned. Not much preprocessing in terms of stemming or lemmatization was possible due to the lack of libraries for this language. We have opted to use the texts as is. We used the same data set for each method. Text Rank requires word embedding, they were generated using Word2Vec. The architecture of twi text sumarization is in Fig 1.

A. TF-IDF score:

TF-IDF stands for Term Frequency - Inverse Document frequency. It is a type of score that quantifies the importance of a word based on its frequency.

$$w_{i,j} = tf_{i,j} * idf_{i,j} \tag{1}$$

1) Term Frequency:: This is a measure of how many times a word is used within a document. The assumption is if a word is repeated several times within a document, it can be assumed that it is related to a central theme in the document. Here, term frequency is calculated as:

$$TermFrequency = 1 + \log(\frac{Keyword\,Count}{Word\,Count}) \tag{2}$$

2) Inverse Document Frequency:: This is a measure of how common or irrelevant a word is. The assumption is if a word is repeated many times across documents it does not hold much topic specific meaning like the word "the". And, inverse document frequency is calculated as :

$$IDF = \log(\frac{No\,of\,Total\,Documents}{No\,of\,Documents\,containing\,the\,term}) \quad (3)$$

For a given document, the TF-IDF score for each word was calculated. The sum of all the scores of words in a sentence was calculated for each sentence, which became the sentence level score. Then the three sentences with the highest score were picked for the summary.

B. Text Rank:

TextRank is an extractive summarization technique, it makes use of the PageRank algorithm to assign scores to each sentence. The PageRank algorithm is used to assign scores to webpages by assessing the importance of a webpage. This is done by counting no. of links and quality of links to a page. A web link from one page to other page is represented as a directed edge. Higher the no. of pages that point to a page higher it's score. A graph is constructed on the basis of these links, and an iterative algorithm is used to continuously update the importance of each node. In PageRank a graph may have nodes that don't lead to any other node.

In TextRank we use the cosine similarity of sentences to construct a link, and it's associated weight. This gives us a weighted edge graph where there is a link between each edge. And we apply the PageRank algorithm on the graph in order to gain scores that reflect the importance of the sentence. The highest scoring sentences are selected for the extractive summary [1].

Embeddings are important to map properties of words in a document. An embedding is a highly dimensional vector that represents the properties of the word in a mathematical form. Embeddings can also be used to compare words, for example the embedding of the word "red" and "apple" will be closer in the vector space. We can consolidate word level embeddings into sentence level embeddings. These sentence vectors can then be used to find the similarity between sentences. In this case, we use it to calculate cosine similarity between sentences. In this work we use Word2Vec in order to generate these embeddings.

C. Fuzzy C-means Clustering:

Fuzzy C-means Clustering is a clustering technique that employs fuzzy logic and set membership. Clusters are defined by the membership of an element's to the fuzzy set of a cluster.

The process consists of defining a Partition matrix that holds the membership values, a set of cluster centres defined by initial membership values, an Updation formula for the membership values based on new cluster centre, and finally a termination condition that defines the point at which we terminate the updation process. This process is done in order to minimize the value of the objective function.

1) Partition Matrix:

$$U = ((\mu_{ij}))_{NxC} \quad (4)$$

Constraints each element is subject to :

$$0 \leq \mu_{ij} \leq 1$$
$$\sum_{j=1}^{c} \mu_{ij} = 1\, for\, all\, i = 1, 2....N$$
$$0 \leq \sum_{j=1}^{c} \mu_{ij} \leq N,\, for\, all\, j = 1, 2...C$$

This matrix holds all the membership values of each and every sentence.

2) Cluster Centre :

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij} \cdot x_i}{\sum_{i=1}^{N} \mu_{ij}} \quad (5)$$

Cluster centres are recalculated for each iteration. The new membership values are measured with respect to the updated centres in each iteration.

3) Membership Updation formula :

$$U_{ij} = \frac{1}{\sum_{k=1}^{C} (\frac{||x_i - c_j||}{||x_i - c_k||})^{\frac{2}{m-1}}} \quad (6)$$

Membership of each element in the matrix is updated using the new cluster centers

4) Objective Function:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} {\mu_{ij}}^{m} ||x_i - c_j||^2 \quad (7)$$

1) Feature Extraction:

1) TF-IDF score : Taken to measure the uniqueness of a sentence. The calculation is same as mentioned above. [5] [7]
2) Numerical Value Score : Sentences that contain numerical information are more likely to contain the significant facts. This is reflected in news articles which contain statistics. Sentences containing figures often represent crucial information

$$NVS = \frac{No.\,of\,Numerical\,Data}{Sentence\,length} \quad (8)$$

3) Sentence Length Score : Used to represent the importance of sentences based on length. Shorter sentences are considered to be less important. Sentences which contain more information are often longer, hence we favour longer sentences with this score.

$$SLS = \frac{Length\,of\,Sentence}{Mean\,Length} \quad (9)$$

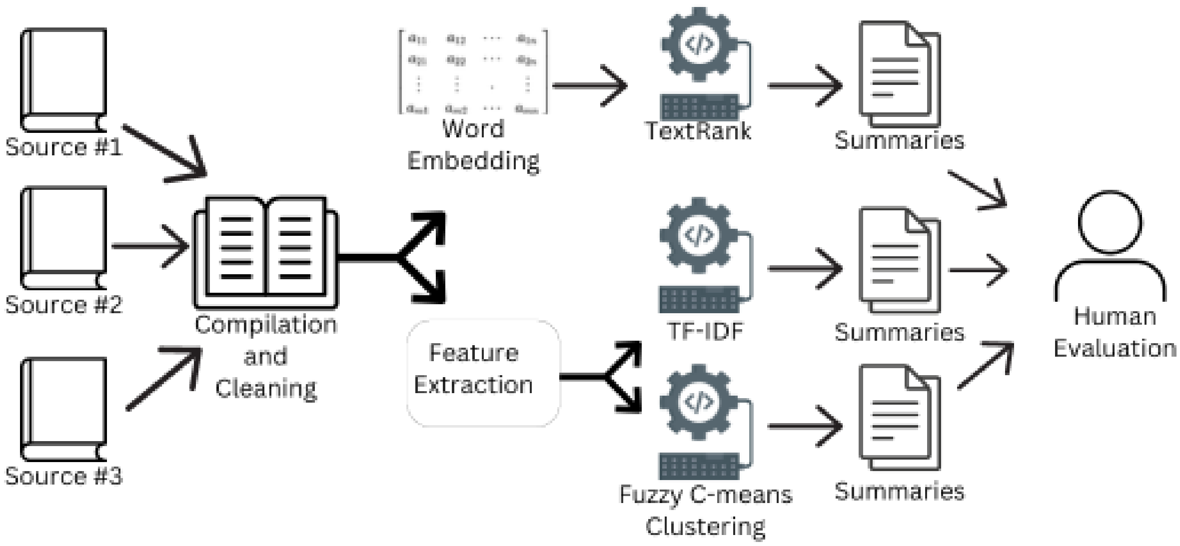

Fig. 1: Architecture of Twi text extractive summarization

4) Positional Scoring : The beginning and ending of a paragraph often contain the more significant information. The first sentence usually expresses the premise of the paragraph and the last sentence delivers the conclusion. Significant amount of information can be conveyed with simply the premise and conclusion.

$$\begin{aligned} If\, Sentence\, Index &< \frac{No.of\, Sentences}{2} \\ PS = \frac{No.of\, Sentences}{2} &- Sentence\, Index \\ If\, Sentence\, Index &> \frac{No.of\, Sentences}{2} \\ PS = Sentence\, Index &- \frac{No.of\, Sentences}{2} \end{aligned} \tag{10}$$

2) Process: The process of generating the Summaries can be summarized as :

1) Features were extracted for each paragraph and stored in individual tables
2) Fuzzy C-means Clustering was employed on the features extracted and clusters were generated
3) The sentence that belonged most to a cluster was chosen for the summary

IV. Results:

We have selected 20 paragraphs from the Twi language and summarized them using the TF-IDF, TextRank, Fuzzy C-means Clustering techniques. We have listed three paragraphs. These summarizations are available in Figure 3. On these summarization, we have conducted a human evaluation for qualitative analysis.

A. Human Evaluation:

Generated summaries were evaluated using human evaluations and each summary was assessed according to the parameters mentioned in Table I. The relevance of the summary, meaningfulness, grammatical correctness and general quality of the summarized text were assessed.

No.	parameters
1	summarised text related to the given? topic
2	main character mentioned?
3	is summarised text meaningful?
4	sentences produced grammatically correct?
5	Overall quality of the summarized text?

TABLE I: Parameters for evaluation

We present the human evaluation scores obtained for summarization using the techniques TF-IDF, TextRank, Fuzzy C-means Clustering in Table II, Table III and Table IV respectively.

Eval	Topic Name	How is the summarized text related to the given topic (%)	Was the name of the main character mentioned in the summarized Text (%)	Are the number of lines summarized meaningful and understandable (%)	Were the sentences produces grammatically correct(%	The overal quality of the summarized text (%)
Eval 1	Para 1	100	80	95	85	90.00
Eval 2	Para 2	100	98	100	100	99.50
Eval 3	Para 3	75	100	100	90	91.25
Eval 4	Para 4	80	100	70	80	82.50
Eval 5	Para 5	90	100	95	80	91.25
Eval 6	Para 6	80	70	90	100	85.00
Eval 7	Para 7	90	90	95	100	95.00
Eval 8	Para 8	80	100	100	100	83.75
Eval 9	Para 9	70	100	90	95	88.75
Eval 10	Para 10	70	83	75	90	79.50
Eval 11	Para 11	100	90	100	70	90.00
Eval 12	Para 12	95	100	100	89	97.25
Eval 13	Para 13	89	100	100	94	95.75
Eval 14	Para 14	99	96	95	83	93.25
Eval 15	Para 15	70	80	79	81	77.50
Eval 16	Para 16	89	68	94	89	85.00
Eval 17	Para 17	74	86	91	100	87.75
Eval 18	Para 18	88	85	77	96	86.50
Eval 19	Para 19	97	95	86	100	94.50
Eval 2 0	Para 20	100	100	90	95	96.25

TABLE II: Results of human evaluation for TF-IDF Summarizations

Eval	Topic Name	How is the summarized text related to the given topic (%)	Was the name of the main character mentioned in the summarized Text (%)	Are the number of lines summarized meaningful and understandable (%)	Were the sentences produces grammatically correct(%	The overall quality of the summarized text (%)
Eval 1	Para 1	100	80	70	100	87.50
Eval 2	Para 2	100	100	90	95	96.25
Eval 3	Para 3	100	100	100	90	97.50
Eval 4	Para 4	100	100	90	80	92.50
Eval 5	Para 5	90	100	90	99	94.75
Eval 6	Para 6	80	70	90	100	85.00
Eval 7	Para 7	90	90	95	100	93.75
Eval 8	Para 8	80	100	100	100	95.00
Eval 9	Para 9	90	100	90	95	93.75
Eval 10	Para 10	70	83	75	90	91.50
Eval 11	Para 11	100	96	100	70	91.50
Eval 12	Para 12	95	100	100	89	96.00
Eval 13	Para 13	89	100	100	94	95.75
Eval 14	Para 14	100	96	95	83	93.50
Eval 15	Para 15	70	80	79	81	77.50
Eval 16	Para 16	90	86	90	89	88.75
Eval 17	Para 17	100	98	91	100	97.25
Eval 18	Para 18	72	79	88	96	83.75
Eval 19	Para 19	91	79	96	100	91.50
Eval 20	Para 20	100	100	90	95	96.25

TABLE III: Results of human evaluation for Text Rank Summarizations

Eval	Topic Name	How is the summarized text related to the given topic (%)	Was the name of the main character mentioned in the summarized Text (%)	Are the number of lines summarized meaningful and understandable (%)	Were the sentences produces grammatically correct(%	The overall quality of the summarized text (%)
Eval 1	Para 1	100	90	100	95	96.25
Eval 2	Para 2	100	70	80	95	86.25
Eval 3	Para 3	85	70	80	90	81.50
Eval 4	Para 4	60	100	60	70	72.50
Eval 5	Para 5	80	95	90	80	87.00
Eval 6	Para 6	70	80	95	100	91.50
Eval 7	Para 7	70	90	95	100	91.25
Eval 8	Para 8	60	100	90	100	83.75
Eval 9	Para 9	60	100	80	99	78.50
Eval 10	Para 10	70	87	75	95	80.00
Eval 11	Para 11	100	100	100	70	92.50
Eval 12	Para 12	95	100	100	89	96.00
Eval 13	Para 13	89	100	100	84	93.25
Eval 14	Para 14	99	56	95	83	83.25
Eval 15	Para 15	100	100	79	81	90.00
Eval 16	Para 16	79	78	94	79	82.50
Eval 17	Para 17	84	66	91	100	85.25
Eval 18	Para 18	86	89	87	99	90.25
Eval 19	Para 19	94	95	88	100	94.25
Eval 20	Para 20	100	100	100	85	96.25

TABLE IV: Results of human evaluation for Fuzzy C-mean Clustering Summarizations

Finally, the Figure 2 shows average percentages obtained from each summarization technique for each parameter.

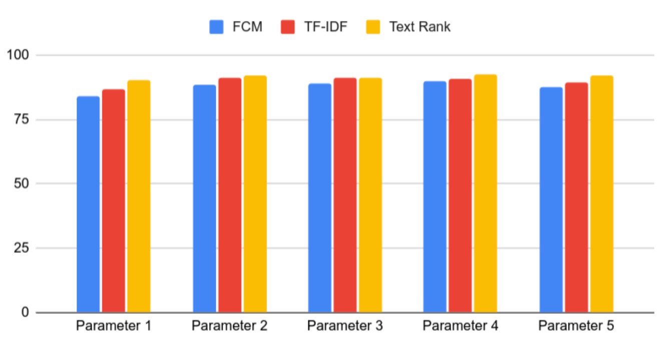

Fig. 2: Results of human evaluation

V. Conclusion

TextRank outperforms the other two methods used. This is believed to be because of the use of embeddings. Word embeddings provide much richer representations. TF-IDF seems to perform comparably well to TextRank, despite it's simplicity, which suggests frequency based approaches are useful with Twi. Fuzzy C-means Clustering performs the worst, which suggests that the features used do not represent the qualities of the text very well. This is a limitation caused by the fact Twi is still a fairly low resource language. However TextRank doesn't seem to be affected by that. This suggests that further research may benefit significantly from approaches based around embeddings, because those approaches are not slowed down because of the researcher's lack of knowledge of the language. Twi extractive summarization has made strides in addressing the challenges specific to the language. Researchers have employed various techniques, utilized available linguistic resources, and explored evaluation metrics. However, further research is needed to overcome existing limitations and advance the state-of-the-art in Twi extractive summarization.

References

[1] Michael Adjeisah, Guohua Liu, Richard Nuetey Nortey, Jinling Song, Khalid Odartey Lamptey, and Felix Nana Frimpong. Twi corpus: a massively twi-to-handful languages parallel bible corpus. In *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 1043–1049. IEEE, 2020.

[2] Satya Dash, Shantipriya Parida, Prosper Mgimwa, Tanjim Taharat Aurpa, and Ahmed Seid. Kiswahili extractive text summarization. 02 2023.

[3] Satya Ranjan Dash, Pubali Guha, Debasish Kumar Mallick, and Shantipriya Parida. Summarizing bengali text: An extractive approach. In *Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, pages 133–140. Springer, 2022.

[4] Muhamad Fahmi Fakhrezi, Moch. Arif Bijaksana, and Arief Fatchul Huda. Implementation of automatic text summarization with textrank method in the development of al-qur'an vocabulary encyclopedia. *Procedia Computer Science*, 179:391–398, 2021. 5th International Conference on Computer Science and Computational Intelligence 2020.

[5] Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva, Fred Freitas, George D.C. Cavalcanti, Rinaldo Lima, Steven J. Simske, and Luciano Favaro. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications*, 40(14):5755–5764, 2013.

[6] Shafiul Gani, Jia Uddin, and Md. Iftekharul Mobin. Extractive text summarization technique using fuzzy c-means clustering algorithm. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, pages 1–6, 2019.

[7] Eduard Hovy. Text summarization. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, Oxford Handbooks in Linguistics, chapter 32, pages 583–598. Oxford University Press, Oxford, 2003.

[8] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics.

[9] Priyanka Pattnaik, Debasish Mallick, Shantipriya Parida, and Satya Dash. *Extractive Odia Text Summarization System: An OCR Based Approach*, pages 136–143. 01 2020.

[10] Joshua Jansen van Vüren and Thomas Niesler. Code-switched language modelling using a code predictive lstm in under-resourced south african languages. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 785–791, 2023.

Para1	Dabi, awarefoɔ bi tenase wɔ kuro bi a yɛ frɛ no dɔ wo yɔnko kurom. Na saa awarefoɔ yi a mereka wɔn ho asɛm yi aware bɛyɛ mfie nsia nso na awoɔ mma. Ne saa nti no, awarefoɔ yi bɛyɛɛ atwetwee maa kuro mma no bi, atɛmdidie ne mpoatwa deɛ wɔ fa no kwa. Ɛdidi mu ara no Ɔdomankoma yɛɛ wɔn adom kyɛɛ awarefoɔ yi Ɔba baa. Na wɔn too ne din sɛ Agudeɛ ɛfiri sɛ Ɔsombo ma na wofoɔ. Afia Agudeɛ yɛ abɔfra bia na ne ho yɛ fɛ paa ara ɔbɔ si . Ne saa nti no biribia ni hɔ a ɔbisa a ne nsa nka. Na n'awofoɔ nsi no fom koraa ɛfiri sɛ wɔn di no ba korɔ. Afei deɛ mfie kakra akyiri no, Afia Agudeɛ na wayini ayɛ fɛfɛɛfɛ yi. Abɔfra yi de ne ahoɔfɛ yi ayini nanso "fofie anto atta" ɛberɛ a Agudeɛ dii mfie bɛyɛ aduonu ɛna ne papa wu gya wɔn. Ɛfiri hɔ no ahokyereɛ kakra baa wɔn akwan mu nanso ɛsiane sɛ Agudeɛ maame pɛ adwuma nti daa na wɔn nsa kɔ wɔn ano.
FCM	Dabi, awarefoÉ" bi tenase wÉ" kuro bi a yÉ› frÉ› no dÉ" wo yÉ"nko kurom.Na wÉ"n too ne din sÉ› AgudeÉ› É›firi sÉ› É"sombo ma na wofoÉ".AbÉ"fra yi de ne ahoÉ"fÉ› yi ayini nanso â€œfofie anto attaâ€ É›berÉ› a AgudeÉ› dii mfie bÉ›yÉ› aduonu É›na ne papa wu gya wÉ"n.
TextRank	Na saa awarefoÉ" yi a mereka wÉ"n ho asÉ›m yi aware bÉ›yÉ› mfie nsia nso na awoÉ" mma.Ne saa nti no, awarefoÉ" yi bÉ›yÉ›É› atwetwee maa kuro mma no bi, atÉ›mdidie ne mpoatwa deÉ› wÉ" fa no kwa.Ne saa nti no biribia ni hÉ" a É"bisa a ne nsa nka.
TF-IDF	Ne saa nti no, awarefoÉ" yi bÉ›yÉ›É› atwetwee maa kuro mma no bi, atÉ›mdidie ne mpoatwa deÉ› wÉ" fa no kwa.AbÉ"fra yi de ne ahoÉ"fÉ› yi ayini nanso â€œfofie anto attaâ€ É›berÉ› a AgudeÉ› dii mfie bÉ›yÉ› aduonu É›na ne papa wu gya wÉ"n.Æ firi hÉ" no ahokyereÉ› kakra baa wÉ"n akwan mu nanso É›siane sÉ› AgudeÉ› maame pÉ› adwuma nti daa na wÉ"n nsa kÉ" wÉ"n ano.
Para2	Ɔnyaa mfeɛ nson no, n'awofoɔ de no kɔɔ abosomfie sɛ ɔnkɔ som ntua ne nana bɛrema bone a w'ayɛ no ka , ɛwɔ "Trojan amamerɛ mu. Wɔn maa no gyidii sɛɛ, sɛ ɔkɔsom wɔ abosomfie hɔ a Ne ho asɛm ɛbɛboa ama ayi ɔhaw ne amaneɛ biaa afiri n'abusuafoɔ so. Ɔsom se abaawa mfie du-nson wɔ abosomfie hɔ wɔ ɛkɔm mu, adwuma dene mu, ɛborɔ mu na afei nso wɔn aamma no ankɔ sukuu, ɔdii mfie du-mmienu no ɔkɔmfo panin a w'adi mfie aduokron too no mɔnaa maa ɔnyinsɛn ne ba a ɛdi kan. Ɔdii mfie aduonu-num no ɔdwane kɔ hyɛɛ dawubɔ a ɛko tia saa ayakayakadeɛ amamerɛ wei a afei abɛyɛ asɛm a ɛda ɛpono so a yɛɛpɛnsɛnpɛnsɛn mu wɔ Ghana man mu. W'ate ankorankorɛ adwumakuo a brɔfo kasa mu no yeferɛ no " non-profit organization" ne "International needs Ghana", a ɛyɛ adwuma yi saa mmaayewa yi firi nkoasom no mu. Ɛnam saa adwumakuo yi so ama mmaayewa bɛɛ apem na anya wɔn faahodie afiri abosomfie bɛɛ du-num mu. Wɔ afe apem aha-nkron aduokron nkron (1999) mu no, ɔgyee "Reebok Human Rights" abasobɔdeɛ.
FCM	WÉ"n maa no gyidii sÉ›É›, sÉ› É"kÉ"som wÉ" abosomfie hÉ" a Ne ho asÉ›m É›bÉ›boa ama ayiÂ É"haw ne amaneÉ› biaa afiri n'abusuafoÉ" so.Æ†som se abaawa mfie du-nson wÉ" abosomfie hÉ" wÉ" É›kÉ"m mu, adwuma dene mu, É›borÉ" mu na afei nso wÉ"n aamma no ankÉ" sukuu, É"dii mfie du-mmienu no É"kÉ"mfo panin a w'adi mfie aduokron too no mÉ"naa maa É"nyinsÉ›n ne ba a É›di kan. Æ†dii mfie aduonu-num no É"dwane kÉ" hyÉ›É› dawubÉ" a É›ko tia saa ayakayakadeÉ› amamerÉ› wei a afei abÉ›yÉ› asÉ›m a É›da É›pono so a yÉ›É›pÉ›nsÉ›npÉ›nsÉ›n mu wÉ" Ghana man mu.W'ate ankorankorÉ› adwumakuo a brÉ"fo kasa mu no yeferÉ› no " non-profit organization" ne "International needs Ghana", a É›yÉ› adwuma yi saa mmaayewa yi firi nkoasom no mu.
TextRank	Æ†som se abaawa mfie du-nson wÉ" abosomfie hÉ" wÉ" É›kÉ"m mu, adwuma dene mu, É›borÉ" mu na afei nso wÉ"n aamma no ankÉ" sukuu, É"dii mfie du-mmienu no É"kÉ"mfo panin a w'adi mfie aduokron too no mÉ"naa maa É"nyinsÉ›n ne ba a É›di kan. Æ†dii mfie aduonu-num no É"dwane kÉ" hyÉ›É› dawubÉ" a É›ko tia saa ayakayakadeÉ› amamerÉ› wei a afei abÉ›yÉ› asÉ›m a É›da É›pono so a yÉ›É›pÉ›nsÉ›npÉ›nsÉ›n mu wÉ" Ghana man mu.W'ate ankorankorÉ› adwumakuo a brÉ"fo kasa mu no yeferÉ› no " non-profit organization" ne "International needs Ghana", a É›yÉ› adwuma yi saa mmaayewa yi firi nkoasom no mu.WÉ" afe apem aha-nkron aduokron nkron (1999) mu no, É"gyee "Reebok Human Rights" abasobÉ"deÉ›.
TF-IDF	Æ†nyaa mfeÉ› nson no, n'awofoÉ" de no kÉ"É" abosomfie sÉ› É"nkÉ" som ntua ne nana bÉ›rema bone a wâ€™ ayÉ› no ka , É›wÉ" "Trojan amamerÉ› mu.WÉ"n maa no gyidii sÉ›É›, sÉ› É"kÉ"som wÉ" abosomfie hÉ" a Ne ho asÉ›m É›bÉ›boa ama ayiÂ É"haw ne amaneÉ› biaa afiri n'abusuafoÉ" so.Æ†som se abaawa mfie du-nson wÉ" abosomfie hÉ" wÉ" É›kÉ"m mu, adwuma dene mu, É›borÉ" mu na afei nso wÉ"n aamma no ankÉ" sukuu, É"dii mfie du-mmienu no É"kÉ"mfo panin a w'adi mfie aduokron too no mÉ"naa maa É"nyinsÉ›n ne ba a É›di kan. Æ†dii mfie aduonu-num no É"dwane kÉ" hyÉ›É› dawubÉ" a É›ko tia saa ayakayakadeÉ› amamerÉ› wei a afei abÉ›yÉ› asÉ›m a É›da É›pono so a yÉ›É›pÉ›nsÉ›npÉ›nsÉ›n mu wÉ" Ghana man mu.
Para3	Asanteman mpaninfoɔ, Enyirisi Aban, Methodist Asɔre ne Presby asɔre na ɛkaa wɔn ho bɔɔmu tee sukuu yi wɔ afe apem aha-nkron aduanan nkron mu (1949). Wɔde sukuu no too Asantehene, Otumfoɔ Osei Tutu Agyeman Prempeh II, a ɔmaa asaase a wɔde sii sukuu no. Wɔde Eton College a ɛwɔ Enyirisiman mu na ɛyɛɛ nhwɛsoɔ. Sukuu yi diikan wɔ nnipa dodoɔ a wɔgyee wɔn wɔ Kwame Nkrumah Suapɔn mu (Kwame Nkrumah University of Science and Technology) wɔ mfeɛ mpem mmienu ne nan (2004) mu wɔ abrɛ a wɔgyee asukuufoɔ 441 na afe 2012 nso, wɔgyee asukuufoɔ 296 firii Prempeh College. Dodoɔ noaa gye tum sɛ sukuu no yɛ ntosoɔ sukuu a edimu wɔ ɔman Ghana mu. Sukuu no adi nkunim mprensa wɔ ɔman mu akansie a w'ato din sɛ "National Robotics Championships" ɛwɔ mpem mmienu ne du-mmiɛnsa (2013) saa kɔsii mfeɛ mpem mmienu ne aduonu baako (2021) mu. Wɔ mfeɛ mpem mmienu ne du-nsia (2016) mu no, Prempeh College nyaa abasobɔdeɛ kɛseɛ a wato din sɛ "Toyota Innovation Award" wɔ amanman akansie a wato din sɛ International Robofest World Championships wɔ Michigan, Amerika man mu
FCM	Sukuu yi diikan wÉ" nnipa dodoÉ" a wÉ"gyee wÉ"n wÉ" Kwame Nkrumah SuapÉ"n mu (Kwame Nkrumah University of Science and Technology) wÉ" mfeÉ› mpem mmienu ne nan (2004) mu wÉ" abrÉ› a wÉ"gyee asukuufoÉ" 441 na afe 2012 nso, wÉ"gyee asukuufoÉ" 296 firii Prempeh College.DodoÉ" noaa gye tum sÉ› sukuu no yÉ› ntosoÉ" sukuu a edimu wÉ" É"man Ghana mu.WÉ" mfeÉ› mpem mmienu ne du-nsia (2016) mu no, Prempeh College nyaa abasobÉ"deÉ› kÉ›seÉ› a wato din sÉ› "Toyota Innovation Award" wÉ" amanman akansie a wato din sÉ› International Robofest World Championships wÉ" Michigan, Amerika man mu
TextRank	Asanteman mpaninfoÉ", Enyirisi Aban, Methodist AsÉ"re ne Presby asÉ"re na É›kaa wÉ"n ho bÉ"É"mu tee sukuu yi wÉ" afe apem aha-nkron aduanan nkron mu (1949).WÉ"de sukuu no too Asantehene, OtumfoÉ" Osei Tutu Agyeman Prempeh II, a É"maa asaase a wÉ"de sii sukuu no.WÉ" mfeÉ› mpem mmienu ne du-nsia (2016) mu no, Prempeh College nyaa abasobÉ"deÉ› kÉ›seÉ› a wato din sÉ› "Toyota Innovation Award" wÉ" amanman akansie a wato din sÉ› International Robofest World Championships wÉ" Michigan, Amerika man mu
TF-IDF	Sukuu yi diikan wÉ" nnipa dodoÉ" a wÉ"gyee wÉ"n wÉ" Kwame Nkrumah SuapÉ"n mu (Kwame Nkrumah University of Science and Technology) wÉ" mfeÉ› mpem mmienu ne nan (2004) mu wÉ" abrÉ› a wÉ"gyee asukuufoÉ" 441 na afe 2012 nso, wÉ"gyee asukuufoÉ" 296 firii Prempeh College.Sukuu no adi nkunim mprensa wÉ" É"man mu akansie a w'ato din sÉ› "National Robotics Championships" É›wÉ" mpem mmienu ne du-mmiÉ›nsa (2013) saa kÉ"sii mfeÉ› mpem mmienu ne aduonu baako (2021) mu.WÉ" mfeÉ› mpem mmienu ne du-nsia (2016) mu no, Prempeh College nyaa abasobÉ"deÉ› kÉ›seÉ› a wato din sÉ› "Toyota Innovation Award" wÉ" amanman akansie a wato din sÉ› International Robofest World Championships wÉ" Michigan, Amerika man mu

Fig. 3: Twi language paragraphs and their summaries using the TF-IDF, TextRank, Fuzzy C-means Clustering