Kiswahili Extractive Text Summarization

Grenze International Journal of Engineering and Technology, Jan Issue

Kiswahili Extractive Text Summarization

Satya Ranjan Dash[1], Shantipriya Parida[2], Prosper Abel Mgimwa[3], Tanjim Taharat Aurpa[4], and Ahmed Iman Seid[5]
[1, 3, 5]KIIT Deemed to be University, India
Email: sdashfca@kiit.ac.in, prospermgimwa1998@gmail.com, Ahmediman14@live.com
[2]SILO AI, Finland
Email: shantipriya.parida@gmail.com
[4]Jahangirnagar University,Bangladesh
Email: taurpa22@gmail.com

***Abstract*—Kiswahili is the language of the Swahili people, who are mostly found in Tanzania and Kenya (along the East African coast and adjacent islands). Although linguistically classified as a Bantu language, Swahili developed historically by borrowing a variety of words from foreign languages, particularly Arabic, around forty percent of the Swahili vocabulary consists of Arabic loan words, including the name of the language an Arabic word meaning ("of the coast"). The number of Swahili speakers is approximated to be around 200 million. In this paper, we have used two Extractive Text Summarization Techniques BERT-SUM and TF-IDF.**

***Index Terms*— Kiswahili Text Summarization, Term Frequency Inverse Document Frequency (TF-IDF), Bidirectional Encoder Representations From Transformers (BERT-SUM), Extractive Text Summarization.**

I. INTRODUCTION

Automatic text summarizing enables people to condense lengthy passages into concise sentences [1]. A language with a long literary history, Kiswahili is also lexically and morphologically abundant. In order to provide individuals with information or knowledge in a highly condensed amount of lines, we have tried to summarize large paragraphs of Swahili by using extractive method of text summarization.

We have applied two techniques on the text. 1) TF-IDF 2) BERT-SUM. Making a long text brief, allows you to appreciate the depth of the long text [2]. This is known as summarizing a text. To summarize a text, one of the many NLP principles that can be applied to many different types of techniques can be employed.

Therefore, there are two distinct approaches to summarize text: the abstractive method and the extractive method. In extractive text summarization, the keywords are taken from the original text and are summarized [3]. This method extracts the keywords without altering the primary source document. In abstractive text summarization, it produces new words and new sentences that provide us with a meaningful summarization like we obtain the summarization from a person. Grammar inconsistencies that exist in extractive approach is solved with the abstractive method.

There are different variations of Extractive Text Summarization techniques. In this paper we have used two methods namely, TF-IDF and BERT-SUM.

Grenze ID: 01.GIJET.9.1.730

II. Related Work

We explored to see if any prior work had been conducted on the Kiswahili language on this subject before deciding to start this research (Extractive Kiswahili Text Summarization). For a language that is spoken widely by a large population this was somehow alarming.

In order to create our own corpus and data set, we have been collecting enormous volumes of Swahili text from numerous sources.

In order to make this feasible for the Kiswahili language, we extended our study on Swahili text summarization. In our research, we found that in-text summarization a machine must first be created as supervised learning. After that, we can incorporate unsupervised learning and determine the response. As a result, it is pre-processed supervised machine learning. Due to the lack of thorough research over the past three decades, Swahili is already structurally and morphologically complex. Therefore, it is extraordinarily difficult for academics and researchers to carry out this kind of study. We have to start from scratch because we are the first researchers to do this research.

III. Experimental Setup

In our research we used 10 paragraphs from a well known Kiswahili novel entitled “JOKA LA MDIMU”, all the paragraphs were summarized using TF-IDF and BERT-SUM techniques and then evaluated manually by five evaluators who understand the language. Below is a figure that shows the manual evaluation procedure:

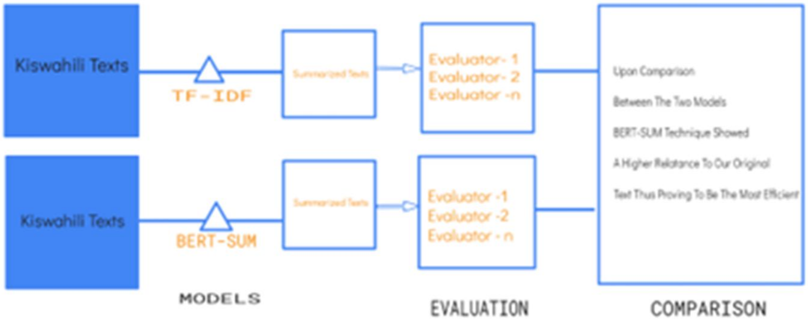

Fig.1 The proposed Model for the Kiswahili Text Summarization Using TF-IDF and Using BERT SUM

A. Original Text Paragraph 1

1) Zamani za miaka kumi sifa kubwa ya Sega ilikuwa uchafu, ulevi, ufuska, na ujambazi. Ilisemwa: Hukatishi Sega usiku. Baadaye wafanyabiashara wakamaizi kuwa idadi kubwa ya wakazi wa Sega wangeweza kuwa washtiri wazuri. Naam, ghafla yakaanza mashindano ya kutafuta nyumba za kufanyia biashara. Kwa vile nyumba za hadhi kama hiyo zilikuwa chache, vibanda vilinunuliwa kwa mamilioni na kubomolewa. Maduka makubwa yakachipua mithili ya uyoga pamoja na majumba ya kulala. Hivi sasa katika barabara kubwa ipitayo katikati ya Sega kila nyumba ni duka, liwe la nguo, nyama, vyombo vya ujenzi na kadhalika. Vichochoro ndani kabisa kuliporomoshwa majumba ya fahari ya kulala, mashine za kupasulia mbao na kusagia nafaka. Viwanda vidogo vikachipuka na kustawi. Ukiachilia hayo, hali ya hapo ikaendelea kuwa duni zaidi ya miaka kumi iliyopita. Kama wenyewe walivyosema: Akheri ya zamani!

B. Corresponding Summarized Text Using BERT-SUM

1) Zamani za miaka kumi sifa kubwa ya Sega ilikuwa uchafu, ulevi, ufuska, na ujambazi. Vichochoro ndani kabisa kuliporomoshwa majumba ya fahari ya kulala, mashine za kupasulia mbao na kusagia nafaka

C. Corresponding Summarized Text Using TF-IDF

1) Ilisemwa: Hukatishi Sega usiku. Viwanda vidogo vikachipuka na kustawi. Kama wenyewe walivyosema: Akheri ya zamani!

IV. Parameters for human evaluation

After getting the summarized texts from each technique ie.TF-IDF and BERT-SUM of every paragraph we held a manual evaluation of the texts using five native speakers of the language to ensure the accuracy of our results, in doing so we had to establish parameters that will guide our evaluation process. The following are the parameters that were taken into account.

TABLE I: THE HUMAN EVALUATION TABLE OF RATING

Parameter No.	**Parameters**
1	How is the summarised text related to the given topic (%)
2	Was the name of the main character mentioned in the summarised Text (%)
3	Are the number of lines summarised meaningful and understandable (%)
4	Were the sentences produced grammatically correct(%)
5	Were the sentences produced grammatically correct(%)

TABLE II. TF-IDF MANUAL EVALUATION

Evaluator	Topic Name (in English)	Is the Summarization is related to the given topic ?	Name of the main character is verified by looking at the Summarization	Presence of the Bag of words is giving a relatable meaning	Is the total no of lines in the Summarization understandable and meaningful ?	Overall quality of the output
Evaluator1	Paragraph 1	100%	90%	57%	57%	Good (79%)
	Paragraph 2	100%	85%	89%	80%	Good (80%)
	Paragraph 3	100%	88%	67%	69%	Good (74%)
	Paragraph 4	100%	82%	56%	56%	Good (82%)
	Paragraph 5	100%	80%	87%	85%	Good (78%)
	Paragraph 6	100%	85%	89%	85%	Good (81%)
	Paragraph 7	100%	90%	78%	75%	Good (70%)
	Paragraph 8	100%	80%	66%	69%	Good (81%)
	Paragraph 9	100%	80%	80%	75%	Good (74%)
Evaluator2	Paragraph 1	100%	93%	67%	69%	Good (83%)
	Paragraph 2	100%	83%	69%	75%	Good (85%)
	Paragraph 3	100%	84%	77%	79%	Good (78%)
	Paragraph 4	100%	80%	56%	58%	Good (76%)
	Paragraph 5	100%	79%	77%	80%	Good (79%)
	Paragraph 6	100%	83%	79%	86%	Good (84%)
	Paragraph 7	100%	89%	77%	74%	Good (80%)
	Paragraph 8	100%	78%	60%	67%	Good (68%)
	Paragraph 9	100%	77%	76%	78%	Good (76%)
Evaluator3	Paragraph 1	100%	90%	59%	67%	Good (69%)
	Paragraph 2	100%	82%	65%	68%	Good (77%)
	Paragraph 3	100%	86%	75%	64%	Good (67%)
	Paragraph 4	100%	79%	64%	56%	Good (77%)
	Paragraph 5	100%	78%	67%	83%	Good (67%)
	Paragraph 6	100%	88%	76%	86%	Good (81%)
	Paragraph 7	100%	94%	86%	78%	Good (78%)
	Paragraph 8	100%	89%	57%	56%	Good (67%)
	Paragraph 9	100%	81%	72%	73%	Good (73%)
Evaluator4	Paragraph 1	100%	94%	70%	69%	Good (80%)
	Paragraph 2	100%	82%	78%	75%	Good (83%)
	Paragraph 3	100%	85%	70%	79%	Good (73%)
	Paragraph 4	100%	87%	65%	58%	Good (65%)
	Paragraph 5	100%	82%	80%	80%	Good (79%)
	Paragraph 6	100%	86%	82%	86%	Good (87%)
	Paragraph 7	100%	94%	76%	74%	Good (76%)
	Paragraph 8	100%	83%	66%	67%	Good (67%)
	Paragraph 9	100%	82%	76%	78%	Good (78%)
Evaluator5	Paragraph 1	100%	91%	63%	68%	Good (78%)
	Paragraph 2	100%	82%	76%	74%	Good (86%)
	Paragraph 3	100%	84%	78%	77%	Good (75%)
	Paragraph 4	100%	86%	68%	59%	Good (67%)
	Paragraph 5	100%	87%	87%	89%	Good (83%)
	Paragraph 6	100%	85%	84%	84%	Good (80%)
	Paragraph 7	100%	94%	72%	73%	Good (73%)
	Paragraph 8	100%	83%	68%	64%	Good (65%)
	Paragraph 9	100%	80%	70%	73%	Good (70%)
	Paragraph 10	100%				

TABLE III. BERT-SUM MANUAL EVALUATION

Evaluator	Topic Name (in English)	Is the Summarization is related to the given topic ?	Name of the main character is verified by looking at the Summarization	Presence of the Bag of words is giving a relatable meaning	Is the total no of lines in the Summarization understandable and meaningful ?	Overall quality of the output
Evaluator1	Paragraph 1	100%	95%	77%	70%	Good (85%)
	Paragraph 2	100%	85%	97%	84%	Good (85%)
	Paragraph 3	100%	70%	87%	76%	Good (83%)
	Paragraph 4	100%	69%	85%	58%	Good (76%)
	Paragraph 5	100%	90%	88%	88%	Good (83%)
	Paragraph 6	100%	88%	80%	89%	Good (84%)
	Paragraph 7	100%	90%	89%	68%	Good (77%)
	Paragraph 8	100%	69%	60%	64%	Good (73%)
	Paragraph 9	100%	72%	76%	80%	Good (79%)
Evaluator2	Paragraph 1	100%	93%	73%	75%	Good (86%)
	Paragraph 2	100%	81%	94%	75%	Good (83%)
	Paragraph 3	100%	66%	86%	79%	Good (80%)
	Paragraph 4	100%	62%	80%	68%	Good (74%)
	Paragraph 5	100%	93%	82%	80%	Good (81%)
	Paragraph 6	100%	85%	84%	76%	Good (83%)
	Paragraph 7	100%	89%	83%	70%	Good (78%)
	Paragraph 8	100%	73%	63%	69%	Good (70%)
	Paragraph 9	100%	69%	78%	70%	Good (76%)
Evaluator3	Paragraph 1	100%	98%	72%	77%	Good (84%)
	Paragraph 2	100%	86%	94%	88%	Good (86%)
	Paragraph 3	100%	60%	88%	54%	Good (76%)
	Paragraph 4	100%	59%	86%	76%	Good (74%)
	Paragraph 5	100%	90%	82%	79%	Good (85%)
	Paragraph 6	100%	83%	76%	85%	Good (83%)
	Paragraph 7	100%	86%	84%	79%	Good (72%)
	Paragraph 8	100%	70%	61%	63%	Good (75%)
	Paragraph 9	100%	67%	75%	70%	Good (74%)
Evaluator4	Paragraph 1	100%	89%	71%	69%	Good (83%)
	Paragraph 2	100%	78%	92%	75%	Good (80%)
	Paragraph 3	100%	64%	83%	79%	Good (81%)
	Paragraph 4	100%	66%	86%	58%	Good (71%)
	Paragraph 5	100%	90%	89%	80%	Good (84%)
	Paragraph 6	100%	82%	83%	86%	Good (85%)
	Paragraph 7	100%	91%	88%	74%	Good (73%)
	Paragraph 8	100%	70%	64%	67%	Good (72%)
	Paragraph 9	100%	62%	69%	78%	Good (69%)
Evaluator5	Paragraph 1	100%	93%	76%	68%	Good (84%)
	Paragraph 2	100%	82%	92%	74%	Good (87%)
	Paragraph 3	100%	65%	86%	77%	Good (86%)
	Paragraph 4	100%	57%	88%	59%	Good (74%)
	Paragraph 5	100%	87%	89%	89%	Good (89%)
	Paragraph 6	100%	82%	78%	84%	Good (89%)
	Paragraph 7	100%	87%	84%	73%	Good (78%)
	Paragraph 8	100%	69%	68%	64%	Good (75%)
	Paragraph 9	100%	61%	74%	73%	Good (78%)

V. METHODOLOGY

Using BERT-SUM (Bidirectional Encoder Representation from Transformers) and TF-IDF (Term Frequency-Inverted Document Frequency) our paragraphs were put into summary .Each paragraph was then broken down in terms of independent sentences and later tokenized into words . Stop-words were employed to cut out phrases that weren't necessary or unique.

After this procedure TF-IDF and BERT-SUM approaches were used corresponding formulas are provided below.

TF = Total Appearance of Word in the Document/Total Words in the Document.

After the TF is calculated, the IDF will be calculated by using the below given formula.

IDF = log (All Document Number/Document Frequency)

TF-IDF = TF*IDF.

Following the completion of the aforementioned calculation, the TDF algorithm arranges the words in the texts in ascending order before determining the rank of each phrase in the document.

VI. RESULT

One can sum up by saying that a workable output was obtained after implying both of the methods to our Swahili paragraphs however the result was not satisfactory due to the limitations of the structure of the Swahili text facilitated by poor or lack of universally accepted, correctly documented stop words, and constraints in the Kiswahili text's framework.

We used human Evaluators to check whether our summarized text had matched the original paragraphs, our evaluators had an excellent understanding, reading and writing knowledge on the Kiswahili Language.

VII. CONCLUSION

The BERT SUM has given better result in the Kiswahili language. In future interactions, we intend to use abstractive tactics to further our study on Kiswahili Text Summarizing Techniques. Other Languages can easily adapt our methods to create summaries for their texts. The Evaluators had to assign percentages of accuracy based on a number of provided parameters see Table 1 and Table 2. In comparison between the two approaches BERT-SUM seemed to be more accurate than TF-IDF approach.

REFERENCES

[1] Pattnaik, P., Mallick, D. K., Parida, S., & Dash, S. R. (2019, December). Extractive odia text summarization system: An ocr based approach. In International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making (pp. 136-143). Springer, Cham.

[2] Dash, S. R., Guha, P., Mallick, D. K., & Parida, S. (2022). Summarizing Bengali Text: An Extractive Approach. In Intelligent Data Engineering and Analytics (pp. 133-140). Springer, Singapore.

[3] Balabantaray, R. C., Sahoo, B., Sahoo, D. K., & Swain, M. (2012). Odia text summarization using stemmer. Int. J. Appl. Inf. Syst, 1(3), 2249-0868.

[4] Abualigah, L., Bashabsheh, M. Q., Alabool, H., & Shehab, M. (2020). Text summarization: a brief review. Recent Advances in NLP: the case of Arabic language, 1-15.

[5] Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., & Affandy, A. (2020). Review of automatic text summarization techniques & methods. Journal of King Saud University-Computer and Information Sciences.

[6] Moratanch, N., & Chitrakala, S. (2017, January). A survey on extractive text summarization. In 2017 international conference on computer, communication and signal processing (ICCCSP) (pp. 1-6). IEEE.

[7] Teklewold, A. A. (2013). Automatic Summarization for Amharic text using open text summarizer (Doctoral dissertation, Addis Ababa University).