IDIAP RESEARCH REPORT

EXTRACTIVE ODIA TEXT SUMMARIZATION SYSTEM: AN OCR BASED APPROACH

Shantipriya Parida[a]

Idiap-RR-02-2020

JANUARY 2020

[a]Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Extractive Odia Text Summarization System: An OCR based Approach

Priyanka Pattnaik[1], Debasish Kumar Mallick[2], Shantipriya Parida[3], and Satya Ranjan Dash[4⋆]

[1] School of Computer Engineering, KIIT University, Odisha, INDIA
{priyankapattanaik2013}@gmail.com
[2] School of Computer Applications, KIIT University, Odisha, INDIA
{mdebasishkumar}@gmail.com
[3] Idiap Research Institute, Martigny Switzerland.
{shantipriya.parida}@idiap.ch
[4] School of Computer Applications, KIIT University, Odisha, INDIA.
{sdashfca}@kiit.ac.in

Abstract. Automatic text summarization is considered as a challenging task in natural language processing field. In the case of multilingual scenario particularly for the low-resource, morphologically complex languages the availability of summarization data set is rare and difficult to construct. In this work, we propose a novel technique to extract Odia text from the image files using optical character recognition (OCR) and summarize the obtained text using extractive summarization techniques. Also, we performed a manual evaluation to measure the quality of summaries to validate our techniques. The proposed approach is found suitable for generating summarized Odia text and the same technique can also extend to other low-resource languages for extractive summarization system.

Keywords: Optical Character Recognition · Text Summarization · Natural Language Processing.

1 Introduction

Automatic text summarization helps human to get their long phrases into short that is full of information and knowledge [2].

Odia is a language which is always known for its literature and it is lexically and morphologically rich. In this paper, we have tried to make a summarization of Odia texts into a few lines so that people can get some information or some knowledge in very few lines [9]. Summarization of text means making a long text in to short so that you can understand the depth of the long text. NLP has so many concepts for a different type of approach and one of its concepts can be done for text summarization. So text summarization can be done by specifically two ways I.e, abstractive way, and extractive way. In extractive text

⋆ Corresponding author

summarization, it extracts the keywords from the source document and gives them a summary [8]. In this approach, the keywords are extracted without making changes in the main source document. In abstractive text summarization, it makes new phrases and new sentences that give us a meaningful summarization like we get the summarization from a human. In abstractive methods, we overcome the grammar inconsistencies that are in extractive method. So in this paper, we have done our work using extractive methods of text summarization.

2 Related Work

When we choose this language for our work, first we look around for the already present corpus for Odia. And we found that the amount is not satisfied for further research and hence there is very less amount of corpus are present with rightly translated in Odia. So it is one of the biggest research gaps for the researchers who have tried to work in this area but could not go for further research. Further, we found that the problem of Odia character recognition is a difficult task for the machine. Due to its high level of design in the character and hence researchers are still currently working on it. To make the machine understand humans, we need to make the machines understand the human's language. And to make this possible for Odia language, we further search for text summarization. In our research, we found that in-text summarization, first we have to make it as supervised learning for the machine to understand and then we can put the unsupervised learning and extract the solution. So it is a preprocessed supervised learning for the machine. In Odia language, its literature is known for its complexity, because it is high in meanings but explained in a single term. So it's a big challenge for all researchers those who are trying to make it easier. While further searching we found that there is no proper stemmer or a lemmatizer for Odia has been made properly. And hence in Odia, it is a difficult part to make a stemmer and a lemmatizer for each word because unlikely in English, Odia has so many extra characters to clean-out from the word and make it the root word. So hence there are so many works to do in Odia and to make it a machine learning language. And hence we found two relevant paper for Odia text summarization. In "Odia Text Summarization Using Stemmer" by R. C. Balabantaray et. al. has done the summarization by using an Odia stemmer [3]. In their work they take a part of the text as for their implementation, first, they had tokenized the whole text and in the second step, they remove the stop words. Then in the third step, they use an Odia stemmer and stem their text, In the fourth step they assigned a weight for words, and the higher rank words come together and make a sentence as for summary. they have done the summary as per the lines user want for them to read i.e, they have done a percentage-wise summary shown to as per user required percentage. S. Biswas et. al. have done an auto text summarization for Odia language [5]. In their work they had used Word frequency method, Positional Criteria method, Cue Phrase method, Title overlap method. By using this method they get the Precision value as and Recall value as 95%. In their work, we found that they had taken fewer stop words and

their work of training is already being trained by the machine so that they got the precision and recall value as 100% for three methods which is used by them [14].

3 Experimental Setup

This section explains the experimental setup used in our paper. The proposed model is shown in Figure 1.

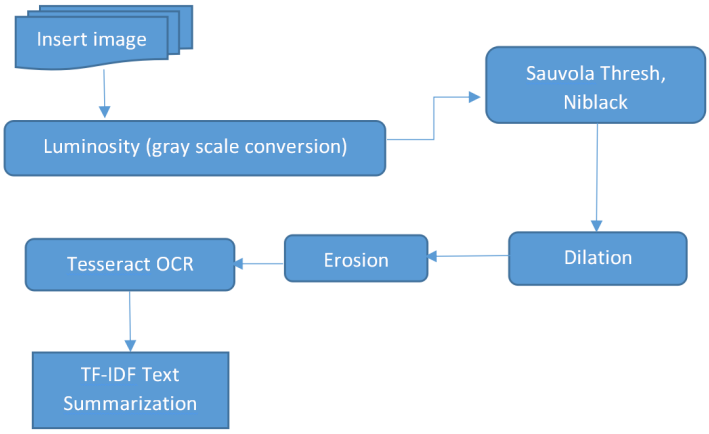


Fig. 1: Proposed Model

3.1 Data Preprocessing

Odia language is rich in its text, hence we take our input as images that may be scanned images or images clicked from any source. The image of Odia written language read as a 3D array. So in our second step, we want the image to be recognized. We have used the "Tesseract OCR engine" to read Odia character [13]. But we found that it uses a traditional method called average method to convert an RGB image to grayscale image. The image is read, in the form of the RGB image. We have converted the RGB image into a grayscale image by applying the weighted method or luminosity method [7]. The formula is given as

$$Grayscale = 0.299R + 0.587G + 0.114B \quad (1)$$

Where R is Red, G is Green and B is blue color in a pixel

As "Tesseract" uses another traditional binarization algorithm called "Otsu" for threshold, so we have used another better algorithm called "Niblack and Sauvola threshold algorithm". And we found that it gives better result as compared to "Otsu" technique. The advantage of Niblack algorithm is that it uses a rectangular window, which slides throughout the image [13]. The center pixel threshold (T) is derived by mean (m) and variance (s) values inside the window.

$$T = m + k*s, \tag{2}$$

Where k is the constant and set that 0.8

But it creates some noise in some area. SO to remove those noises we have included Sauvola algorithm. Thus the modified formula is

$$T = m*(1-k*(1-(s/R))), \tag{3}$$

Where R is the dynamics of standard deviation that is set to 128

This formula will not detect all the images of documents. So the normalized formula we have implemented is

$$T = m-k*(1-(s/R))*(m-M)), \tag{4}$$

Where R is the maximum standard deviation of all the windows, M is the gray level of the image

However, some pixels are missed during these processes which may lead to an error of character recognition. So we have used Dilation which helps to join those areas where some pixel values are missed [6].

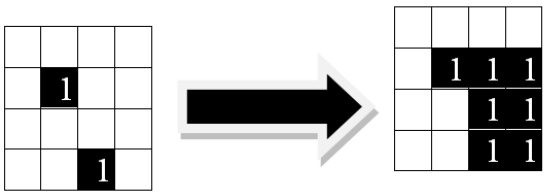


Fig. 2: Dilation

We can see that in the left diagram (Figure 2) some pixels are missed. And we can see that when the "Dilation" algorithm is implemented as shown in the right side (Figure 2) it also have a problem, i.e. it takes extra pixels. For reducing this kind of errors the "Erosion" algorithm is implemented as shown in Figure 3.

In our preprocessed data we take four Odia images i.e, the images related to "Naveen Pattnaik", "Debasish Mohanty", "Narendra Modi", and "Subash Chandra Bose". The sample image containing Odia text for one of the topic "Debasish Mohanty" is shown in Figure 4.

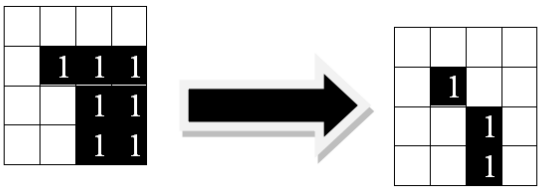
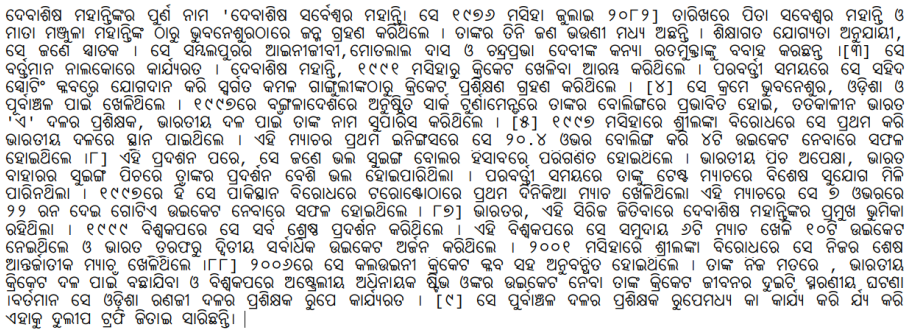


Fig. 3: Erosion



ଦେବାଶିଷ ମହାନ୍ତିଙ୍କର ପୂର୍ଣ ନାମ 'ଦେବାଶିଷ ସର୍ବେଶ୍ୱର ମହାନ୍ତି' ସେ ୧୯୭୬ ମସିହା ଜୁଲାଇ ୨୦[୨] ତାରିଖରେ ପିତା ସବେଶ୍ୱର ମହାନ୍ତି ଓ ମାତା ମଞ୍ଜୁଳା ମହାନ୍ତିଙ୍କ ଠାରୁ ଭୁବନେଶ୍ୱରଠାରେ ଜନ୍ମ ଗ୍ରହଣ କରିଥିଲେ । ତାଙ୍କର ତିନି ଜଣ ଭଉଣୀ ମଧ୍ୟ ଅଛନ୍ତି । ଶିକ୍ଷାଗତ ଯୋଗ୍ୟତା ଅନୁଯାୟୀ, ସେ ଜଣେ ସ୍ନାତକ । ସେ ସମ୍ବଲପୁରର ଆଇନୀଙ୍ଗୀବୀ, ମୋତଲାଲ ଦାସ ଓ ଚନ୍ଦ୍ରପ୍ରଭା ଦେବୀଙ୍କ କନ୍ୟା ରତମୁଛାଙ୍କୁ ବବାହ କରଛନ୍ତୁ ।[୩] ସେ ବର୍ତ୍ତମାନ ନାଲକୋରେ କାର୍ଯ୍ୟରତ । ଦେବାଶିଷ ମହାନ୍ତି, ୧୯୯୧ ମସିହାରୁ କ୍ରିକେଟ ଖେଳିବା ଆରମ୍ଭ କରିଥିଲେ । ପରବର୍ତ୍ତୀ ସମୟରେ ସେ ସହିଦ ଖୋଟିଂ କ୍ଲବରୁ ଯୋଗଦାନ କରି ସ୍ୱର୍ଗତ କମଳ ଗାଙ୍ଗୁଲୀଙ୍କଠାରୁ କ୍ରିକେଟ ପ୍ରଶିକ୍ଷଣ ଗ୍ରହଣ କରିଥିଲେ । [୪] ସେ କ୍ରମେ ଭୁବନେଶ୍ୱର, ଓଡ଼ିଶା ଓ ପୂର୍ବାଞ୍ଚଳ ପାଇଁ ଖେଳିଥିଲେ । ୧୯୯୭ରେ ବଙ୍ଗଳାଦେଶରେ ଅନୁଷ୍ଠିତ ସାର୍କ ଟୁର୍ଣାମେନ୍ଟରେ ତାଙ୍କର ବୋଲିଙ୍ଗରେ ପ୍ରଭାବିତ ହୋଇ, ତତକାଳୀନ ଭାରତ 'ଏ' ଦଳର ପ୍ରଶିକ୍ଷକ, ଭାରତୀୟ ଦଳ ପାଇଁ ତାଙ୍କ ନାମ ସୁପାରିସ କରିଥିଲେ । [୫] ୧୯୯୭ ମସିହାରେ ଶ୍ରୀଲଙ୍କା ବିରୋଧରେ ସେ ପ୍ରଥମ କରି ଭାରତୀୟ ଦଳରେ ସ୍ଥାନ ପାଇଥିଲେ । ଏହି ମ୍ୟାଚର ପ୍ରଥମ ଇନିଙ୍ଗସରେ ସେ ୨୦.୪ ଓଭର ବୋଲିଙ୍ଗ କରି ୪ଟି ଉଇକେଟ ନେବାରେ ସଫଳ ହୋଇଥିଲେ ।[୮] ଏହି ପ୍ରଦଶନ ପରେ, ସେ ଜଣେ ଭଲ ସୁଇଙ୍ଗ ବୋଲର ହସାବରେ ପରଗଣତ ହୋଇଥିଲେ । ଭାରତୀୟ ପଚ ଅପେକ୍ଷା, ଭାରତ ବାହାରର ସୁଇଙ୍ଗ ପିଚରେ ତାଙ୍କର ପ୍ରଦର୍ଶନ ବେଶି ଭଲ ହୋଇପାରିଥିଲା । ପରବର୍ତ୍ତୀ ସମୟରେ ତାଙ୍କୁ ଟେଷ୍ଟ ମ୍ୟାଚରେ ବିଶେଷ ସୁଯୋଗ ମିଳି ପାରିନଥିଲା । ୧୯୯୭ରେ ହିଁ ସେ ପାକିସ୍ତାନ ବିରୋଧରେ ଟରୋଣ୍ଟୋଠାରେ ପ୍ରଥମ ଦିନିକିଆ ମ୍ୟାଚ ଖେଳିଥିଲେ। ଏହି ମ୍ୟାଚରେ ସେ ୭ ଓଭରରେ ୨୨ ରନ ଦେଇ ଗୋଟିଏ ଉଇକେଟ ନେବାରେ ସଫଳ ହୋଇଥିଲେ । [୭] ଭାରତର, ଏହି ସିରିଜ ଜିତିବାରେ ଦେବାଶିଷ ମହାନ୍ତିଙ୍କର ପ୍ରମୁଖ ଭୁମିକା ରହିଥିଲା । ୧୯୯୯ ବିଶ୍ୱକପରେ ସେ ସର୍ବ ଶ୍ରେଷ୍ଠ ପ୍ରଦର୍ଶନ କରିଥିଲେ । ଏହି ବିଶ୍ୱକପରେ ସେ ସମୁଦାୟ ୬ଟି ମ୍ୟାଚ ଖେଳି ୧୦ଟି ଉଇକେଟ ନେଇଥିଲେ ଓ ଭାରତ ତରଫରୁ ଦ୍ୱିତୀୟ ସର୍ବାଧିକ ଉଇକେଟ ଅର୍ଜନ କରିଥିଲେ । ୨୦୦୧ ମସିହାରେ ଶ୍ରୀଲଙ୍କା ବିରୋଧରେ ସେ ନିଜର ଶେଷ ଆନ୍ତର୍ଜାତୀକ ମ୍ୟାଚ ଖେଳିଥିଲେ ।[୮] ୨୦୦୬ରେ ସେ କଲଉଇନୀ କ୍ରିକେଟ କ୍ଲବ ସହ ଅନୁବନ୍ଧୁତ ହୋଇଥିଲେ । ତାଙ୍କ ନିଜ ମତରେ , ଭାରତୀୟ କ୍ରିକେଟ ଦଳ ପାଇଁ ବଛାଯିବା ଓ ବିଶ୍ୱକପରେ ଅଷ୍ଟ୍ରେଲୀୟ ଅଧିନାୟକ ଷ୍ଟିଭ ଓଙ୍କର ଉଇକେଟ ନେବା ତାଙ୍କ କ୍ରିକେଟ ଜୀବନର ଦୁଇଟି ସ୍ମରଣୀୟ ଘଟଣା ।ବର୍ତମାନ ସେ ଓଡ଼ିଶା ରଣଜୀ ଦଳର ପ୍ରଶିକ୍ଷକ ରୁପେ କାର୍ଯ୍ୟରତ । [୯] ସେ ପୂର୍ବାଞ୍ଚଳ ଦଳର ପ୍ରଶିକ୍ଷକ ରୁପେମଧ୍ୟ କା କାର୍ଯ୍ୟ କରି ଯ୍ୟ କରି ଏହାକୁ ଦୁଲୀପ ଟ୍ରଫି ଜିତାଇ ସାରିଛନ୍ତି। |

Fig. 4: An example of original image containing Odia text

Table 1: Manual evaluation parameters for the generated summaries.

Parameter	Description
Parameter1	Is the summarization is related to the given topic ?
Parameter2	Name of the main character is verified by looking at the summarization
Parameter3	Presence of the Bag of words is giving a relatable meaning
Parameter4	Is the total no of lines in the summarization understandable and meaningful ?
Parameter5	Overall quality of the output

3.2 Methodology

After given the perfect shape, the “Tesseract” tool kit performs Odia character extraction. For text summarization, we have used “Term Frequency-Inverse Document Frequency”. The sentences which are extracted from the image are tokenized which split them into sentences. After sentences are tokenized, the sentences are split into words. To remove unnecessary words which are present in the sentences, the stop-word filtration process is performed [10,4]. As in Odia language, less number of a stop-word dataset is present. We have made our dataset.

After removing of stop-words, the rest of the words “Term-Frequency (TF)” are calculated by the given formula below

Extracted Data

ଦେବାଶିଷ ମହାନ୍ତିଙ୍କର ପୁର୍ଣ ନାମ 'ଦେବାଶିଷ ସର୍ବେଶ୍ୱର ମହାନ୍ତି । ସେ ୧୯୭୭ ମସିହା ଜୁଲାଇ ୨୦[୨] ତାରିଖରେ ପିତା ସବେଶ୍ୱର ମହାନ୍ତି ଓ ମାତା ମଞ୍ଜୁଲା ମହାନ୍ତିଙ୍କ ଠାରୁ ଭୁବନେଶ୍ୱରଠାରେ ଜନ୍ମ ଗ୍ରହଣ କରିଥିଲେ । ତାଙ୍କର ତିନି ଜଣ ଭଉଣୀ ମଧ୍ୟ ଅଛନ୍ତି । ଶିକ୍ଷାଗତ ଯୋଗ୍ୟତା ଅନୁଯାୟୀ, ସେ ଜଣେ ସ୍ୱପାତକ । ସେ ସମ୍ବଲପୁରର ଆଇନୀଜୀବୀ,ମୋତିଲାଲ ଦାସ ଓ ଚନ୍ଦ୍ରପ୍ରଭା ଦେବୀଙ୍କ କନ୍ୟା ରିତିମୁକ୍ତାଙ୍କୁ ବିବାହ କରିଛନ୍ତି ।[୩] ସେ ବର୍ତ୍ତମାନ ନାଲକୋରେ କାର୍ଯ୍ୟରତ । ଦେବାଶିଷ ମହାନ୍ତି, ୧୯୯୧ ମସିହାରୁ କ୍ରିକେଟ ଖେଳିବା ଆରମ୍ଭ କରିଥିଲେ । ପରବର୍ତ୍ତୀ ସମୟରେ ସେ ସହିଦ ସ୍ୱପୋର୍ଟିଂ କ୍ଲବରେ ଯୋଗଦାନ କରି ସ୍ୱର୍ଗତ କମାଳ ଗାଙ୍ଗୁଲୀଙ୍କଠାରୁ କ୍ରିକେଟ ପ୍ରଶିକ୍ଷଣ ଗ୍ରହଣ କରିଥିଲେ । [୪] ସେ କ୍ରମେ ଭୁବନେଶ୍ୱର, ଓଡ଼ିଶା ଓ ପୂର୍ବାଞ୍ଚଳ ପାଇଁ ଖେଳିଥିଲେ । ୧୯୯୭ରେ ବଙ୍ଗାଲାଦେଶରେ ଅନୁଷ୍ଠିତ ସାର୍କ ଟୁର୍ଣାମେନ୍ଟରେ ତାଙ୍କର ବୋଲିଙ୍ଗରେ ପ୍ରଭାବିତ ହୋଇ, ତତକାଳୀନ ଭାରତ 'ଏ' ଦଳର ପ୍ରଶିକ୍ଷକ, ଭାରତୀୟ ଦଳ ପାଇଁ ତାଙ୍କ ନାମ ସୁପାରିସ କରିଥିଲେ । [୫] ୧୯୯୭ ମସିହାରେ ଶ୍ରୀଲଙ୍କା ବିରୋଧରେ ସେ ପ୍ରଥମ କରି ଭାରତୀୟ ଦଳରେ ସ୍ଥାନ ପାଇଥିଲେ । ଏହି ମ୍ୟାଚର ପ୍ରଥମ ଇନିଙ୍ଗସରେ ସେ ୨୦.୪ ଓଭର ବୋଲିଙ୍ଗ କରି ୪ଟି ଉଇକେଟ ନେବାରେ ସଫଳ ହୋଇଥିଲେ ।[୮] ଏହି ପ୍ରଦର୍ଶନ ପରେ, ସେ ଜଣେ ଭଲ ସୁଇଙ୍ଗ ବୋଲର ହିସାବରେ ପରିଗଣିତ ହୋଇଥିଲେ । ଭାରତୀୟ ପିଚ ଅପେକ୍ଷା, ଭାରତ ବାହାରର ସୁଇଙ୍ଗ ପିଚରେ ତାଙ୍କର ପ୍ରଦର୍ଶନ ବେଶି ଭଲ ହୋଇପାରିଥିଲା । ପରବର୍ତ୍ତୀ ସମୟରେ ତାଙ୍କୁ ଟେଷ୍ଟ ମ୍ୟାଚରେ ବିଶେଷ ସୁଯୋଗ ମିଳି ପାରିନଥିଲା । ୧୯୯୭ରେ ହିଁ ସେ ପାକିସ୍ଥାନ ବିରୋଧରେ ଟରୋଣ୍ଟୋଠାରେ ପ୍ରଥମ ଦିନିକିଆ ମ୍ୟାଚ ଖେଳିଥିଲେ। ଏହି ମ୍ୟାଚରେ ସେ ୭ ଓଭରରେ ୨୨ ରନ ଦେଇ ଗୋଟିଏ ଉଇକେଟ ନେବାରେ ସଫଳ ହୋଇଥିଲେ । [୭] ଭାରତର, ଏହି ସିରିଜ ଜିତିବାରେ ଦେବାଶିଷ ମହାନ୍ତିଙ୍କର ପ୍ରମୁଖ ଭୁମିକା ରହିଥିଲା । ୧୯୯୯ ବିଶ୍ୱକପରେ ସେ ସର୍ବ ଶ୍ରେଷ୍ଠ ପ୍ରଦର୍ଶନ କରିଥିଲେ । ଏହି ବିଶ୍ୱକପରେ ସେ ସମୁଦାୟ ୬ଟି ମ୍ୟାଚ ଖେଳି ୧୦ଟି ଉଇକେଟ ନେଇଥିଲେ ଓ ଭାରତ ତରଫରୁ ଦ୍ୱିତୀୟ ସର୍ବାଧିକ ଉଇକେଟ ଅର୍ଜନ କରିଥିଲେ । ୨୦୦୧ ମସିହାରେ ଶ୍ରୀଲଙ୍କା ବିରୋଧରେ ସେ ନିଜର ଶେଷ ଆନ୍ତର୍ଜାତୀକ ମ୍ୟାଚ ଖେଳିଥିଲେ ।[୮] ୨୦୦୬ରେ ସେ କଲଉଇନୀ କ୍ରିକେଟ କ୍ଲବ ସହ ଅନୁବନ୍ଧିତ ହୋଇଥିଲେ । ତାଙ୍କ ନିଜ ମତରେ , ଭାରତୀୟ କ୍ରିକେଟ ଦଳ ପାଇଁ ବଛାଯିବା ଓ ବିଶ୍ୱକପରେ ଅଷ୍ଟ୍ରେଲୀୟ ଅଧିନାୟକ ଷ୍ଟିଭ ତାଙ୍କର ଉଇକେଟ ନେବା ତାଙ୍କ କ୍ରିକେଟ ଜୀବନର ଦୁଇଟି ସ୍ମରଣୀୟ ଘଟଣା ।ବର୍ତମାନ ସେ ଓଡ଼ିଶା ରଣଜୀ ଦଳର ପ୍ରଶିକ୍ଷକ ରୁପେ କାର୍ଯ୍ୟରତ । [୯] ସେ ପୁର୍ବାଞ୍ଚଳ ଦଳର ପ୍ରଶିକ୍ଷକ ରୁପେ ମଧ୍ୟ କା କାର୍ଯ୍ୟ କରି ର୍ଯ୍ୟ କରି ଏହାକୁ ଦୁଲୀପ ଟ୍ରଫି ଜିତାଇ ସାରିଛନ୍ତି।

Summarized Data

ସେ ୧୯୭୭ ମସିହା ଜୁଲାଇ ୨୦[୨] ତାରିଖରେ ପିତା ସବେଶ୍ୱର ମହାନ୍ତି ଓ ମାତା ମଞ୍ଜୁଲା ମହାନ୍ତିଙ୍କ ଠାରୁ ଭୁବନେଶ୍ୱରଠାରେ ଜନ୍ମ ଗ୍ରହଣ କରିଥିଲେ ଦେବାଶିଷ ମହାନ୍ତି, ୧୯୯୧ ମସିହାରୁ କ୍ରିକେଟ ଖେଳିବା ଆରମ୍ଭ କରିଥିଲେ ୧୯୯୭ରେ ବଙ୍ଗାଲାଦେଶରେ ଅନୁଷ୍ଠିତ ସାର୍କ ଟୁର୍ଣାମେନ୍ଟରେ ତାଙ୍କର ବୋଲିଙ୍ଗରେ ପ୍ରଭାବିତ ହୋଇ, ତତକାଳୀନ ଭାରତ 'ଏ' ଦଳର ପ୍ରଶିକ୍ଷକ, ଭାରତୀୟ ଦଳ ପାଇଁ ତାଙ୍କ ନାମ ସୁପାରିସ କରିଥିଲେ ଏହି ମ୍ୟାଚର ପ୍ରଥମ ଇନିଙ୍ଗସରେ ସେ ୨୦.୪ ଓଭର ବୋଲିଙ୍ଗ କରି ୪ଟି ଉଇକେଟ ନେବାରେ ସଫଳ ହୋଇଥିଲେ ୧୯୯୭ରେ ହିଁ ସେ ପାକିସ୍ଥାନ ବିରୋଧରେ ଟରୋଣ୍ଟୋଠାରେ ପ୍ରଥମ ଦିନିକିଆ ମ୍ୟାଚ ଖେଳିଥିଲେ। ଏହି ମ୍ୟାଚରେ ସେ ୭ ଓଭରରେ ୨୨ ରନ ଦେଇ ଗୋଟିଏ ଉଇକେଟ ନେବାରେ ସଫଳ ହୋଇଥିଲେ ତାଙ୍କ ନିଜ ମତରେ , ଭାରତୀୟ କ୍ରିକେଟ ଦଳ ପାଇଁ ବଛାଯିବା ଓ ବିଶ୍ୱକପରେ ଅଷ୍ଟ୍ରେଲୀୟ ଅଧିନାୟକ ଷ୍ଟିଭ ତାଙ୍କର ଉଇକେଟ ନେବା ତାଙ୍କ କ୍ରିକେଟ ଜୀବନର ଦୁଇଟି ସ୍ମରଣୀୟ ଘଟଣା [୯] ସେ ପୁର୍ବାଞ୍ଚଳ ଦଳର ପ୍ରଶିକ୍ଷକ ରୁପେ ମଧ୍ୟ କା କାର୍ଯ୍ୟ କରି ର୍ଯ୍ୟ କରି ଏହାକୁ ଦୁଲୀପ ଟ୍ରଫି ଜିତାଇ ସାରିଛନ୍ତି

Fig. 5: An example of extracted Odia text and generated summaries

TF = Total Appearance of Word in the Document/Total Words in the Document
After calculating the Term-Frequency, the Inverse Document Frequency (IDF) will calculate. The Formula is given below

IDF = log (All Document Number/Document Frequency)
TF-IDF = TF*IDF

After been calculated, words of the documents are sorted in descending order by its TF-IDF. By summation of all TF-IDF values of words present in the sentences, which decide the rank of sentences values [1,12]. As TF-IDF is an

Table 2: Human evaluation rating table.

Evalutor	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	Naveen Pattnaik	100%	80%	55%	55%	Good (75%)
	Debasish Mohanty	100%	88%	55%	65%	Good (80%)
	Narendra Modi	100%	80%	60%	60%	Good (76%)
	Subash Chandra Bose	100%	85%	75%	65%	Good (84%)
Evaluator2	Naveen Pattnaik	100%	80%	65%	75%	Good (63%)
	Debasish Mohanty	100%	90%	60%	60%	Good (75%)
	Narendra Modi	100%	90%	55%	60%	Good (72%)
	Subash Chandra Bose	100%	85%	63%	64%	Good (85%)
Evaluator3	Naveen Pattnaik	100%	70%	68%	65%	Good (67%)
	Debasish Mohanty	100%	80%	66%	67%	Good (74%)
	Narendra Modi	100%	75%	57%	59%	Good (69 %)
	Subash Chandra Bose	100%	85%	85%	84%	Good (80%)
Evaluator4	Naveen Pattnaik	100%	80%	69%	67%	Good (66%)
	Debasish Mohanty	100%	90%	66%	69%	Good (72%)
	Narendra Modi	100%	60%	60%	60%	Good (62%)
	Subash Chandra Bose	100%	95%	85%	85%	Good (80%)

extractive method the sentences appear in after summarization, those are same as the sentences present in the original document [11].

4 Result

When the proposed technique applied to the selected data, we got the summarized text as per our desire. The extracted Odia text and the generated Odia summaries are shown in Figure 5.

To judge the summarization, we have evaluated our results by human evaluators. We have chosen four human evaluators who can read, write and understands Odia properly. We have set five parameters for the manual evaluation as mentioned in the Table 1.

We have decided to do the human evaluation as in our case we find it difficult for automatic evaluation. So, we provide the Odia extracted text and the generated summaries to the four experts (person who know Odia, who can write Odia properly, who can read Odia properly, and who can understand Odia properly). According to their evaluation, we find that all our result are purely related to the extracted Odia text hence the summarization is related to the topic. According to our result, they also have gone through our evaluation criteria and they give us result in percentile format. The manual evaluation results are shown in the Table 2.

5 Conclusion and Future Work

In this paper, we have proposed a method for extracting Odia text from the image and generating summarized text. Odia is a language which is rich in text and known for its literature but lack in computational resources for machines to perform different NLP tasks such as machine translation, summrization, etc. Our motive is to make Odia language more enhanced for the machines by creating more language resources. In our future work, we will consider the abstractive techniques for summarizing Odia text by building summarization dataset (Odia texts and its corresponding summaries). Our method can be easily extended to generate summaries for low resource language.

References

1. Aizawa, A.: An information-theoretic perspective of tf–idf measures. Information Processing & Management **39**(1), 45–65 (2003)
2. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268 (2017)
3. Balabantaray, R., Sahoo, B., Sahoo, D., Swain, M.: Odia text summarization using stemmer. Int. J. Appl. Inf. Syst **1**(3), 2249–0868 (2012)
4. Bharti, S.K., Babu, K.S.: Automatic keyword extraction for text summarization: A survey. arXiv preprint arXiv:1704.03242 (2017)
5. Biswas, S., Acharya, S., Dash, S.: Automatic text summarization for oriya language. International Journal of Computer Applications **975**, 8887 (2015)
6. Gaikwad, D.K., Mahender, C.N.: A review paper on text summarization. International Journal of Advanced Research in Computer and Communication Engineering **5**(3), 154–160 (2016)

7. Joshi, N.: Text image extraction and summarization. Asian Journal For Convergence In Technology (AJCT) (2019)
8. Kryściński, W., Paulus, R., Xiong, C., Socher, R.: Improving abstraction in text summarization. arXiv preprint arXiv:1808.07913 (2018)
9. Lloret, E.: Text summarization: an overview. Paper supported by the Spanish Government under the project TEXT-MESS (TIN2006-15265-C06-01) (2008)
10. Munot, N., Govilkar, S.S.: Comparative study of text summarization methods. International Journal of Computer Applications **102**(12) (2014)
11. Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al.: Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023 (2016)
12. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 133–142. Piscataway, NJ (2003)
13. Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
14. Yousefi-Azar, M., Hamey, L.: Text summarization using unsupervised deep learning. Expert Systems with Applications **68**, 93–105 (2017)