

Regional Language Text based Enhancement of Summarization Factuality using Machine Learning

Ms. Apurva D. Dhawale¹ and Dr. Sonali B. Kulkarni²

¹⁻²Dept of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Ch. Sambhajinagar, Maharashtra, India
Email: addhawale@gmail.com, sonalibkul@gmail.com

Abstract—The Text summarization of any lengthy text is opportune need for everyone now a days. It gives us fine, substantive short summary out of the large text and saves time of an individual. So, it is a trending technique for practical usage by people in each area and sector. There are 2 categories of text summarization: 1. Abstractive and 2. Extractive. This study focuses on Extractive text summarization using unsupervised TextRank algorithm approach. Here, we are concentrating on the regional language of Maharashtra i.e. Marathi as it proves to be the unnoticed one while studying text summarization techniques in several domains.

Index Terms—Abstractive Summarization, Extractive summarization, NLP, Marathi text, machine learning, unsupervised.

I. INTRODUCTION

The Ministry of Education of India, Section 29(f) of Chapter V under Right to Education Act, 2009 clearly states that, “medium of instructions shall, as far as practicable, be in child’s mother tongue” [1], Hence the regional languages captivated vital importance in the field of education.

There is also a lot of large and descriptive content on the internet, such as media and other educational resources as a result, there is a need for an accessible and functional text summarizing technique that can acquire important content from a huge quantity of text while reducing the individual personal effort in obtaining relevant knowledge. Researchers have put a lot of effort into developing a text summarizing platform that can generate summaries for any text. However, discrepancies and poor precision have plagued these methods, making them unsuitable for this purpose. [2]

Hence, this study focuses on Marathi language, we are trying to present anew approach for text summarization of Marathi news articles. Here we considered news articles from 5 different domains like Sports, Polity, Banking, General Knowledge and Entertainment/film industry. Here we have considered a Marathi news articles dataset of 1135 files in .txt format. We have applied Unsupervised TextRank machine learning algorithm on single document and performed extractive text summarization.

II. LITERATURE SURVEY

Anishka Chaudhari proposed a system for extractive text summarization method using neural networks for Marathi text. Extractive summaries or extracts are produced by identifying important sentences or words which are directly selected from the document.

To perform extractive text summarization we propose to use a Recurrent Neural Network (RNN) – a type of neural network that can perform calculations on sequential data (e.g. sequences of words) – as it has become the standard approach for many Natural Language Processing tasks. The translation of the Marathi text to English will be done using the Google translate API for this proposed system. [3]

Document summarization technique helps to overcome the obstacle available in the document generation [4–9]. The research work [10–12] takes input sentence from the document, and preprocesses it by separating words, removing stop words, and then the last step is stemming.

Merged model graph is used to analyze the summaries in the single paragraph and that is compared with the evaluation summary model [12,13].

Furthermore, N-gram graph Powered Evaluation via Regression (NPowER) is the most important scheme and that evaluates the machine learning issues and provides the optimal solution for the evaluation strategies [14].

III. PROPOSED METHODOLOGY

The TextRank algorithm supports Gensim library which is best suitable for Marathi text summarization. Gensim library has 3 function supported for text summary generation.

The first is Summary() which is a default function to give 0.2 i.e. 20% summary of the original text.

The second is wordcount(), which takes parameters, and we have considered summary generation using 150 words.

The third one is by using Ratio(), which has default value of 0.2, but it can be changed. So, we considered 50% i.e. 0.5 (value) summary of the original text.

All these methods are applied after preprocessing of input text. The domains specified are all involved in training and testing phase. Here, 60% articles are used for training the data and 40% are used to test. We have 250 news articles of Sports, Polity, Banking, and film industry domains each in the available dataset. Here we considered 100 articles for testing. The General Knowledge domain has total 135 articles in the dataset, here we considered 54 articles for testing.

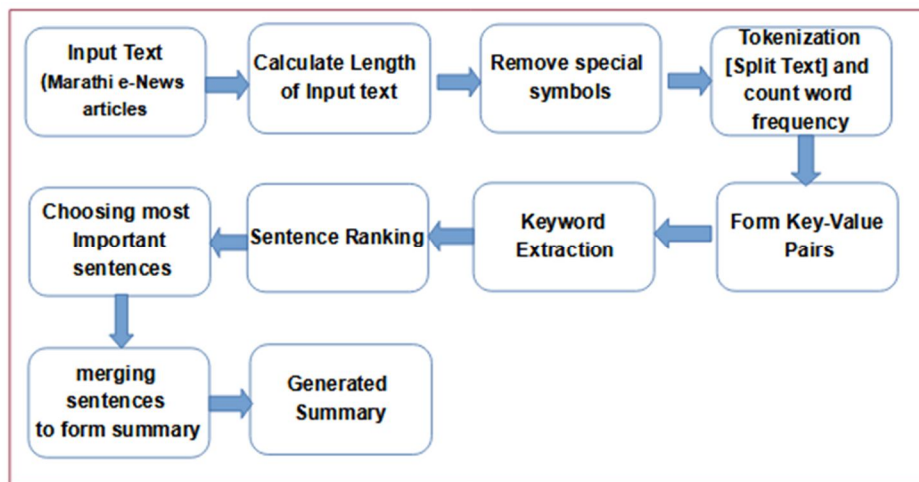


Figure 1. Proposed Methodology

Here, the preprocessing steps are specified which includes calculating total length, tokenization, removing special symbols, count the frequency of words, form key-value pairs, extract keywords. Every step gives us a structured data or text from input text.

Gensim is an open source library which means “generate similar”. After preprocessing the text the sentences are ranked and the most important sentences are chosen for summary generation.

IV. EXPERIMENTAL ANALYSIS AND RESULT EVALUATION

As stated above, the Gensim library functions are applied on all the domains and observed results are shown below:

The first domain is Banking. We applied all 3 methods on 100 samples each and determined the following outcome:

TABLE I: BANKING NEWS SUMMARIZATION EVALUATION USING SUMMARY () = 0.2, RATIO = 0.5 & WORDCOUNT= 150 WORDS

Banking News Summarization Evaluation					
Summary () = 0.2, Ratio = 0.5 & Wordcount= 150 words					
Sr. No	News	Total Length	Summary	Summary by Ratio	Summary by Wordcount
1	b1	2372	460	1381	929
2	b2	2141	641	1366	1118
3	b3	1843	372	936	1028
4	b4	1431	336	912	1091
5	b5	562	149	347	561
6	b6	1425	426	762	945
7	b7	1302	395	723	991
8	b8	1423	467	869	1080
9	b9	660	117	300	659
10	b10	2016	537	1034	1034

This is the sample result of Banking News Summarization Evaluation where there is observation of 8 positives and two negative result. The summary which was not up to the mark was listed in negatives. We obtained this result for 100 samples and its graphical representation is shown below:

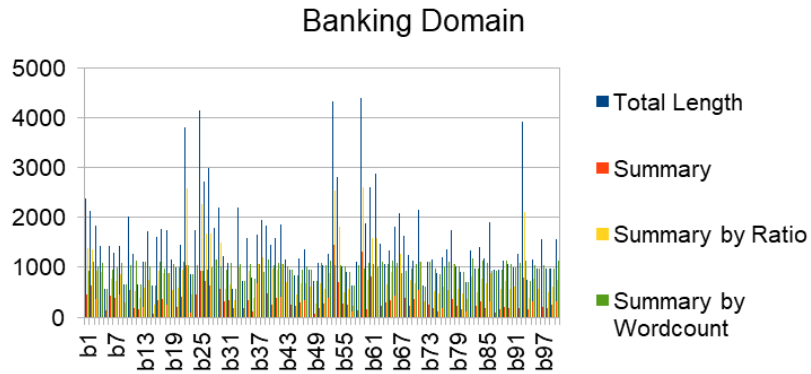


Figure 2: Graphical Result of text summarization for banking domain

The second domain is Sports. Here the observations are like below:

TABLE II: SPORTS NEWS SUMMARIZATION EVALUATION USING SUMMARY () = 0.2, RATIO = 0.5 & WORDCOUNT= 150 WORDS

Sports News Summarization Evaluation					
Summary()=0.2, Ratio = 0.5 & Wordcount= 150 words					
Sr. No	News	Total Length	Summary	Summary by Ratio	Summary by Wordcount
1	S1	1715	253	793	1025
2	S2	2105	577	1183	1005
3	S3	2829	830	1817	1070
4	S4	1361	259	735	1049
5	S5	775	141	464	773
6	S6	1431	327	1014	1049
7	S7	1202	233	673	1092
8	S8	904	225	536	898
9	S9	545	82	373	544
10	S10	706	140	433	705

This is the sample result of sports News Summarization Evaluation where there is observation of 7 positives and 3 negative result. The summary which was not up to the mark was listed in negatives. Graphical representation:

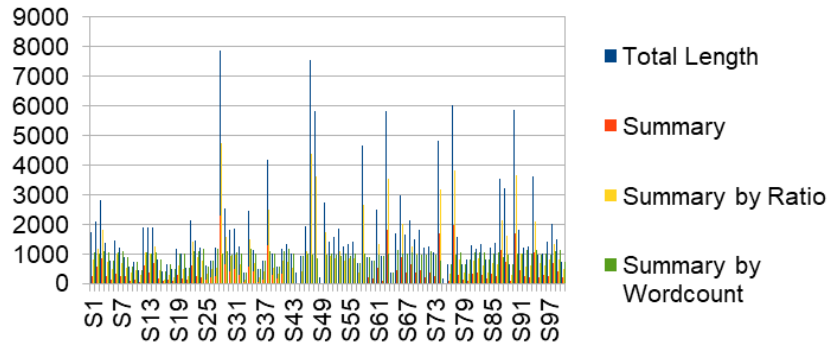


Figure 3: Graphical Result of text summarization for sports domain

The third domain is film industry. Here, the results we got are shown below:

TABLE III: FILM INDUSTRY NEWS SUMMARIZATION EVALUATION USING SUMMARY ()=0.2, RATIO = 0.5 & WORDCOUNT= 150 WORDS

Film Industry News Summarization Evaluation					
Summary()=0.2, Ratio = 0.5 & Wordcount= 150 words					
Sr. No	News	Total Length	Summary	Summary by Ratio	Summary by Wordcount
1	n1	1780	304	1202	1081
2	n2	899	231	597	892
3	n3	557	125	288	556
4	n4	854	280	519	847
5	n5	1788	529	966	882
6	n6	189	0	34	188
7	n7	3198	725	1724	943
8	n8	1292	372	638	1040
9	n9	2049	663	1292	1049
10	n10	796	171	354	791

This is the sample result of Film Industry News Summarization Evaluation where there is observation of 7 positives and 3 negative result. The summary which was not up to the mark was listed in negatives. Graphical representation:

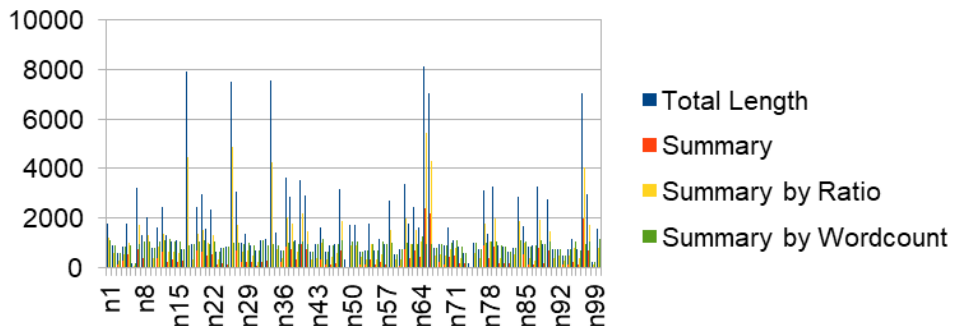


Figure 4: Graphical Result of text summarization for Film Industry domain

The fourth domain is General Knowledge. The observed results are:

TABLE IV: GENERAL KNOWLEDGE NEWS SUMMARIZATION EVALUATION USING SUMMARY() $\approx$ 0.2, RATIO = 0.5 & WORDCOUNT= 150 WORDS

General knowledge News Summarization Evaluation					
Summary() $\approx$ 0.2, Ratio = 0.5 & Wordcount= 150 words					
Sr. No	News	Total Length	Summary	Summary by Ratio	Summary by Wordcount
1	doc1	2157	514	1145	1079
2	doc2	3531	832	2028	985
3	doc3	2518	658	1523	1024
4	doc4	1123	286	736	1120
5	doc5	1021	137	610	1017
6	doc6	2446	639	1484	1020
7	doc7	1579	433	1010	1093
8	doc8	908	213	466	904
9	doc9	1204	204	613	1105
10	doc10	664	167	396	662

This is the sample result of General Knowledge News Summarization Evaluation where there is observation of 9 positives and 1 negative result. The summary which was not up to the mark was listed in negatives. Graphical representation:

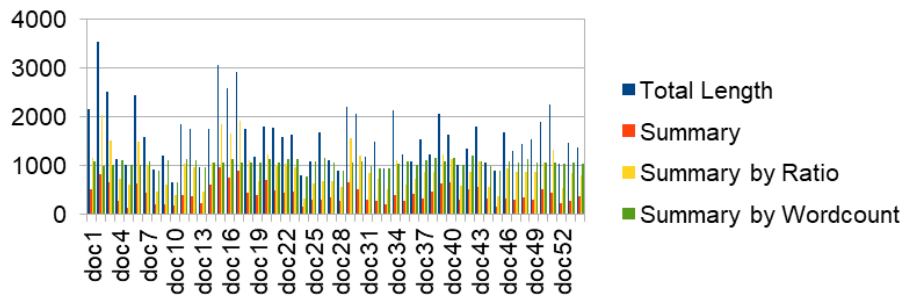


Figure 5: Graphical Result of text summarization for General Knowledge domain

The last domain is Polity. Results we got are as:

TABLE V: POLITY NEWS SUMMARIZATION EVALUATION USING SUMMARY() $\approx$ 0.2, RATIO = 0.5 & WORDCOUNT= 150 WORDS

Polity News Summarization Evaluation					
Ratio = 0.5 & Wordcount= 150 words					
Sr. No	News	Total Length	Summary	Summary by Ratio	Summary by Wordcount
1	P1	2107	653	1343	1037
2	P2	1351	384	870	1096
3	P3	1161	180	738	1035
4	P4	705	146	382	704
5	P5	775	146	508	775
6	P6	4052	1358	2831	996
7	P7	725	189	414	719
8	P8	1017	247	523	1017
9	P9	803	118	425	802
10	P10	628	161	415	628

This is the sample result of Polity News Summarization Evaluation where there is observation of 5 positives and 5 negatives result. The summary which was not up to the mark was listed in negatives. Graphical representation:

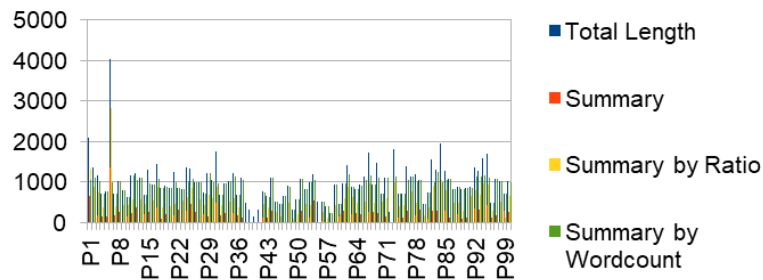


Figure 6: Graphical Result of text summarization for Polity domain

The obtained collective results are as:

TABLE VI: RESULTS

Domain	Total	positives	Negatives	Percentage	Average
Banking	100	77	23	77.00	81.2
Sports	100	84	16	84.00	
Film Industry	100	77	23	77.00	
General Knowledge	54	49	5	91.00	
Polity	100	77	23	77.00	

The observed results are then plotted and its graphical representation is as below:

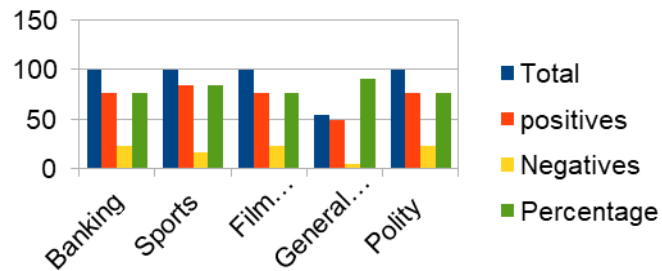


Figure 7: Combined Graphical Result of text summarization for all domains considered

This graph and tables shows the results of all domains which are calculated on the basis of the comparative study values of positive and negative summary samples. So the overall average accuracy of all the methods combined is 81.2%. This shows that the system really works well with Marathi text unsupervised extractive summary generation.

V. CONCLUSION

The paper theorize the research study which is performed on the Marathi e-news of 5 different domains like Banking, Sports, Film Industry, General Knowledge, & Politics. The Unsupervised TextRank algorithm is used for summarizing News articles and the outcomes shows that it works well with General Knowledge and sports with 91% and 84% respectively. Result for all other domains are observed to be 77% respectively. Considering the positives and negatives in the news samples, the results of all three methods are compared and the average accuracy is 81.2 %. This summarization method may establish to be more meticulous and accommodating to the society.

REFERENCES

- [1] <https://Www.Pib.Gov.In/Pressreleasepage.aspx?RID=1847061#:~:Text=As20envisaged%20in%20para%204.11,%2FIocal%20language%2Fregional%20language.>

- [2] Sunil D. Kale*, Parikshit N. Mahalle, Renu Kachhoria, Santosh Kumar, Prasad Chaudhari, Vivek D. Patil, "Marathi Text Summarization Through NLP And Deep Learning Mechanism ", Journal Of Autonomous Intelligence, Volume 6 Issue 3 Doi: 10.32629/Jai.V6i3.1009, (2023).
- [3] Anishka Chaudhari¹, Akash Dole², Deepali Kadam³, "Marathi Text Summarization Using Neural Networks ", International Journal Of Advance Research And Development, (Volume 4, Issue 11) 2019.
- [4] Liu, X., Webster, J.J., Kit, C.: An Extractive Text Summarizer Based On Significant Words. In: Iccpol 2009 Lnai, Vol. 5459, Pp. 168–178. Springer, Berlin/Heidelberg (2009)
- [5] García-Hernández, R.A., Montiel, R., Ledeneva, Y., Rendón, E., Gelbukh, A., Cruz, R.: Text Summarization By Sentence Extraction Using Unsupervised Learning. In: Gelbukh, A., Morales, E.F. (Eds.) Micai 2008 Lncs (Lnai), Vol. 5317, Pp. 133–143. Springer, Heidelberg (2008)
- [6] Luhn, H.P.: The Automatic Creation Of Literature Abstracts. Ibm J. Res. Dev. 2(2), 159–165 (1958)
- [7] Edmondson, H.P.: New Methods In Automatic Extraction. J. Assoc. Comput. Mach. 16(2), 264– 285 (1969)
- [8] Goldstein, J., Carbolell, J., Kantrowitz, M., Mittal, V.: Summarizing Text Documents: Sentence And Evaluation Metrics. In: 22nd Int. Acm Sigir Research And Development In Information Retrieval, Berkley (1999) A Survey Of Distinctive Prominence Of Automatic Text Summarization. . . 549
- [9] Ledeneva, Y., Hernández, R.G., Soto, R.M., Reyes, R.C., Gelbukh, A.: Emclustering Algorithm For Automatic Text Summarization. In: Batyrshin, I., Sidorov, G. (Eds.) Micai 2011, Part I Lnai, Vol. 7094, Pp. 305–315. Springer, Berlin/Heidelberg (2011)
- [10] [Http://Www.Cdacnoida.in/Snlp/Digital_Library/Text_Summ.Asp](http://www.cdacnoida.in/Snlp/Digital_Library/Text_Summ.Asp)
- [11] Gupta, V., Lehal, G.S.: Automatic Punjabi Text Extractive Summarization System. In: International Conference On Computational Linguistics, Coling 2012, Pp. 191–198. Iit, Bombay (2012)
- [12] Babar, S.A., Patil, P.D.: Improving Performance Of Text Summarization. In: International Conference On Information And Communication Technologies, Procedia Computer Science, Vol. 46, Pp. 354–363. Elsevier (2014)
- [13] Al Qassem, L.M., Wang, D., Al Mahmoud, Z., Barada, H., Al-Rubaie, A., Almoosa, N.I.: Automatic Arabic Summarization: A Survey Of Methodologies And Systems. In: 3rd International Conference On Arabic Computational Linguistics, Acling, 5–6 Nov 2017. Elsevier, Dubai
- [14] Giannakopoulos, G.: Multi-Document Multilingual Summarization And Evaluation Tracks In Acl 2013 Multiling Workshop. In: Proceedings Of The Multiling 2013 Workshop On Multilingual Multidocument Summarization, Pp. 20–28 (2013)