**Advances and Applications in Mathematical Sciences**
Volume 20, Issue 12, 2021, Pages 3351-3363


# A HYBRID APPROACH FOR EXTRACTIVE DOCUMENT SUMMARIZATION WITH BIG DATA ANALYTICS

**V. VADIVU and N. KAVITHA**

Department of Computer Science
Nehru Arts and Science College
Coimbatore, India

## Abstract

With the experience of the web, there is an immense expansion in the capacity of data. It is an assortment of monstrous and complex data sets that incorporate tremendous amounts of data, social media analytics, data the executive's capabilities, ongoing data, and so forth. Looking through applicable data in an assortment of archives is a monotonous assignment. The arrangement that comes into the picture for this issue is programmed text summarization. In this paper, the rundown age of enormous archives of archives for big data is proposed, considering client contribution as a topic. Big Data Analytics refers to methods used to examine and secure knowledge from big data. Hence, big data analytics can see as a sub-measure in the general cycle of 'understanding extraction' from big data. The consequence of applying to cluster improves the summarizer framework to gather actual words instead of duplicating excess words utilizing the WordNet Tool. Topic-based summarization from big data is a testing task, especially when various archives have the same or diverse substance. With its programming methods, Hadoop can give better ways of creating rundown, and it likewise improves the multifaceted nature of summarization measure utilizing disseminated Computing.

## 1. Introduction

As data keeps developing violently, we want to rapidly and precisely get the data required from the gigantic data measures, such as the web, big data stage. Even though it is a big test, archive summarization can give a rundown of the first reports to fulfill the need for short comprehension for related substance. Consequently, programmed multi-record summarization for the data assets turns into an examination centre [3]. Big data alludes to dynamic data produced in complex frameworks with the three Vs' attributes:

2010 Mathematics Subject Classification: 68.
Keywords: Big data, Map Reduce, Summarization, LDA, Word Net Tool, *K*-medoids.
Received November 20, 2020; Accepted December 19, 2020

volume, velocity, and assortment [2]. Topic models are different levelled probabilistic models that have their beginnings in the field of AI. Topic models have been completely applied, especially in composing assessment [4] free data sources with passed on and decentralized controls are a principal attribute of Big Data applications. Acting naturally, each data source can create and assemble data without including (or relying upon) any joined control [7]. Extraction type summary is to directly select phrase or sentences of high importance from the original document and combined them [8].

Natural Language Processing community has been investigating the domain of summarization for nearly the last half-century [9].

Text summarization procedures subsequently produce a smaller abstract of lone or various chronicles. The accompanying framework passes on the crucial data in the first text(s). Applications fuse coherent and reports, advertisements, messages, and destinations. Completely, summarization follows two strategies: the extractive cycle and the abstractive process. In an extractive diagram, a framework is produced using the leading content units (usually sentences). The following overview is a subset of the preliminary report. Strangely, abstractive summarization techniques incorporate isolating semantic data from the content.

Modified watchword extraction is the route toward picking words and articulations from the content document that can, most ideal situation, adventure the record's middle is feeling with no human mediation depending upon the model [1]. The goal of modified expression extraction is to utilize the power and speed of current count abilities to access and recuperate, pushing upon data relationship without human annotators' extra costs. Summarization is where the most remarkable features of a book are isolated and joined into the preliminary report [2].

LDA Model is a probabilistic generative topic model. It can locate the most recent topics in the data set. Through LDA (Latent Dirichlet Allocation) Model [7], a grouping of records is isolated into some latent issues; that is, each report is viewed as a mix scattering of topics probability movement where the weight of a topic addresses topic-importance for that chronicle. Furthermore, these latent topics are displayed as mixes of words with a probability allotment, saying the terms' scores for each issue.

The rest of this paper has coordinated as follows. Section 2 presents the related work about document summarization. The proposed work has introduced in section 3. The results and discussions are implemented in section 4. The conclusion is delivered in section 5.

## 2. Related Work

K. T. Belerao, et al. [1] proposed another structure to make an exciting layout from records' immense social affair using the MapReduce framework. The Implementation utilizes an open-source Java library for getting semantic similarity words and recalling a definitive target to discover any topic data from a bunch of points of reference from the gigantic data determined graph is made. It will spare clients from investigating through each record. All hypothetical at one will be open. J. Bian, et al. [3] because of the LDA Model, the topic-dissemination of documents and the term-allocation of topics are gotten. In the paper, as shown by the generative sentence cycle, determining topic-criticalness, and the topic-dissemination of sentences, we propose another sentence-situating technique to get the fantastic nature of sentences. R. K. Lomotey, et al. [5] with the current corporate trades, it is clear that "Big Data" has come to remain. This is because most stock business trades that utilization to be paper-based are, all in all, being digitized.

Furthermore, the customer delivered content across the different scope of the endeavour scene is extending in volume at an exceptional rate. While Big Data has its enormous ideal conditions, the data is heterogeneous (i.e., collection) presents new challenges. Xindong Wu et al. [7] explored a couple of troubles at the data, model, and structure levels. To help Big Data mining, tip-top handling stages are required, constraining systematic designs to deliver the Big Data's full power. At the data level, the independent data sources and the combination of the data grouping conditions often achieve data with tangled prerequisites, for instance, missing/flawed characteristics. In various conditions, insurance concerns, disturbance, and errors can be brought into the data to make altered data copies. Developing a liberated from any damage data sharing show is a critical test. The key test is to create overall models at the model level by joining secretly discovered guides to outline a coupling together view.

## 3. System Model

*K*-means clustering is a procedure for vector quantization, at first from signal readiness, known for some data mining assessments. *K*-means clustering expects to allocate recognitions into $k$ packs in which each observation has a spot with the gathering with the nearest mean, filling in as a model of the part. The groups achieve isolating the data space into Voronoi cells. The issue is computationally problematic (NP-hard); regardless, there are beneficial heuristic algorithms normally used and meet quickly to a close-by ideal. These are ordinarily like the craving support algorithm for Gaussian movements' blends using an iterative refinement approach utilized by the two algorithms. Additionally, both of them use bunch centers around demonstrating the data. Regardless, $k$-means clustering will find gatherings of indistinguishable spatial degrees for all intents and purposes, while the craving extension segment licenses packs to have different shapes.

The $k$-medoids algorithm is a clustering algorithm related to the $k$-means algorithm and the medoid-move algorithm. Both the $k$-means and $k$-medoids algorithms are partitional (breaking the dataset up into social affairs). $k$-medoid is a conventional allotting clustering method that bundles the data set of $n$ objects into $k$ gatherings known from before. An important gadget for concluding $k$ is the layout. It is more generous to clatter and special cases when appeared differently concerning $k$-means. It restricts the number of pairwise dissimilarities instead of a measure of squared Euclidean detachments. A medoid can be described as the object of a pack whose ordinary uniqueness to all the gathering articles is irrelevant. For instance, it is the most mostly discovered point in the gathering.

The drawback of the $k$-means algorithm in multi archive summarization is that it takes a lot of time to cluster the keywords created in the third module. In contrast, the $k$-medoids algorithm diminishes the time in clustering the keywords in the third module. There will be exceptions in the k-means algorithm with the end goal that while clustering, it might miss some data that lessen the precision of the summarization. As there are no exceptions in the $k$-medoids algorithm, the archive's summarization will be more precise than utilizing the $k$-means algorithm.

Summarization is performed through four stages. They are

- Document Clustering
- Latent Dirichlet Algorithm (LDA)
- Semantic Similar Terms Generation and clustering utilizing *k*-medoids
- Sentence Filtering

The dataset separated from Case history from the Federal court of Australia is utilized as the dataset. It contains the document history in XML records. Clustering and topic modeling is applied to the dataset to produce a synopsis. At first, the data under the text tag in the XML record is separated and contributed to the primary module. In the primary module, comparable reports will be assembled under a bunch utilizing the text clustering strategy. Clustering is the way toward gathering comparable archives. Text clustering bunches comparable archives under a group utilizing the document names. LDA is utilized during the time spent on topic modeling.

Topic modeling is the way toward allocating topic terms to record bunches. Topic terms are the words that often happen in the report. Before producing topic terms, stop word disposal will take place utilizing POS labelling. POS tagger give labels to each word present in the archives. Stop words alludes to the most well-known words utilized in a language. Generally, articles and relational words are viewed as stop words. The stop words will have their labels, and utilizing the labels; the stop words will get killed. After producing the topic terms, relative semantic terms to these terms are created utilizing WordNet.

WordNet is an API that is equipped for creating similar words for the given the word. Word weightage for these terms is determined utilizing the TF-IDF method. This procedure gives weight to each word, and the words are bunched utilizing the *K*-medoids clustering algorithm. *K*-medoids is an apportioning algorithm. The sentences containing the topic terms and their particular semantic comparative terms are sifted and stuck in the summed up archive. The yield record will give a surmised rundown of the assortment of archives.

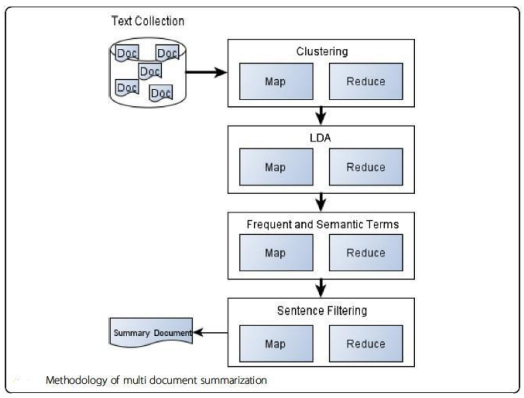


**Figure 1.** Methodology of multi-document summarization.

In the block diagram, the archive assortment goes through a clustering measure. Comparative records will be assembled under each bunch. LDA produces the topic term for each group utilizing the topic modeling procedure. The comparative semantic terms for topic terms are created in the following stage. The created terms are bunched utilizing the $k$-medoids algorithm. The words with the highest weightage are picked for taking an interest in the outline archive.

### 3.1. Document Clustering

Clustering is the way toward getting sorted out items into bunches whose individuals are comparable somehow or another. Clustering of reports is finished by gathering the archives dependent on the names of XML records. The text clustering method is applied to the multi report assortment to make the archive bunches. Similar text reports can be gathered under their particular groups for making them prepared for summarization along these lines.

In this module, at first, the client needs to transfer the dataset. The records present in the dataset will be assembled into bunches dependent on the document names. The client needs to choose the group to be summed up, and that bunch will be passed to the next module.

### 3.2. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a topic modeling procedure applied to every individual text archive present in a group. At the first stop word, the

disposal cycle will take place utilizing the concept of POS labeling. POS labeling is the procedure that gives labels to all the words in the archive, dependent on their properties. All the stop words will have their labels, which can be utilized as a key to killing them. After taking out stop words, recurrence of the leftover words will be determined, and afterwards utilizing LDA topic terms will be produced. Topic terms are the words that happen much of the time in a cluster.

### 3.3. Frequent and Semantic Similar Terms Generation

The cluster topics produced from the past module will be given as a contribution to this module. Semantic comparable terms for the topic terms are created utilizing WordNet. WordNet is an API that can produce comparative words for a given the word. Semantic terms allude to the words which have the same significance as the chose word. The frequencies of the comparative semantic terms are determined. Word weightage count is finished by utilizing the TF-IDF idea. TF-IDF is a mathematical measurement technique that is planned to reflect how important a word is to a record in an assortment. It is utilized as a weighting factor in data recovery and text mining. After ascertaining the loads, the clustering of words is finished. During the time spent clustering the words, the $K$-medoids clustering algorithm is utilized. $K$-medoids is a parcelling procedure that clusters the data set of $n$ objects into $k$ clusters.

In this module, the semantic terms will be given weightage by giving the often happened words score. The words which have a score of more than 0.3 will be clustered utilizing the $k$-medoids algorithm.

### 3.4. Sentence Filtering

Sentence filtering is performed from every individual info text report present in the archive cluster. The words which are clustered in the past stage are given as a contribution to this module. Clusters which are having words with more weightage are given the highest need. The words present in the clusters with the most noteworthy need are chosen for sentence filtering measure. The sentences which are containing these words are picked and stuck in the outline archive.

### 3.5. *K*-Medoid Clustering Algorithm

*K*-Medoid clustering is also a partition-based clustering algorithm. It uses medoids to represent the clusters. A medoid represents the most centrally located data item of the data set. In medoid, the data member of a data set whose average dissimilarity to all the other members of the set is minimal.

Input: number of clusters $k$, the data set containing $n$ items $D$.

Output: A set of $k$ clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoids.

$$z = \sum_{i=k}^{k} \sum | x - mi |$$

Where, $Z$ is Sum of absolute error for all items in the data set, $x$ is the data point in the space representing a data item and $m_i$ is the medoid of cluster $C_i$ Process:

1. Arbitrarily choose $k$ data items as the initial medoids.

2. Assign each remaining data item to a cluster with the nearest medoid.

3. Randomly select a non-medoid data item and compute the total cost of swapping old medoid data item with the currently selected non-medoid data item.

4. If the total cost of swapping is less than zero, then perform the swap operation to generate the new set of $k$-medoids.

5. Repeat steps 2, 3 and 4 till the medoids stabilize their locations.

## 4. Results and Discussion

The Proposed system has implemented Java Programming Language with a Hadoop environment.

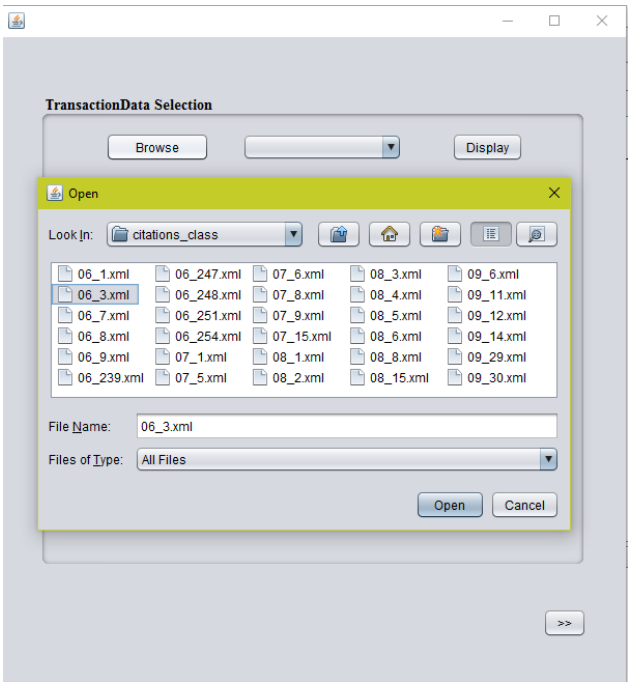


**Figure 2.** Dataset Selection.

Figure 2 shows the documents in the dataset to be uploaded.

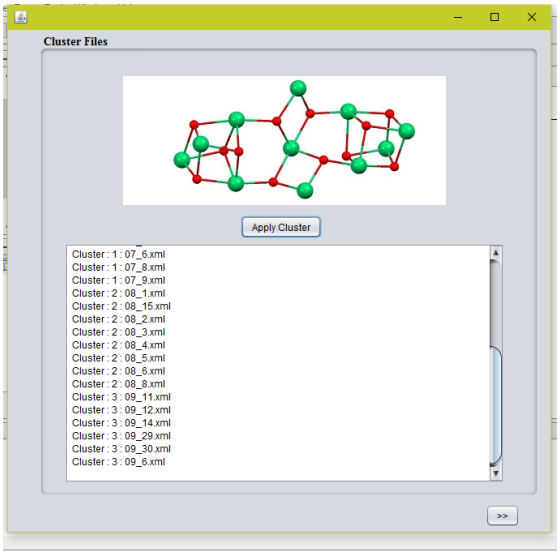


**Figure 3.** Applying Cluster.

Figure 3 shows the documents under each cluster.

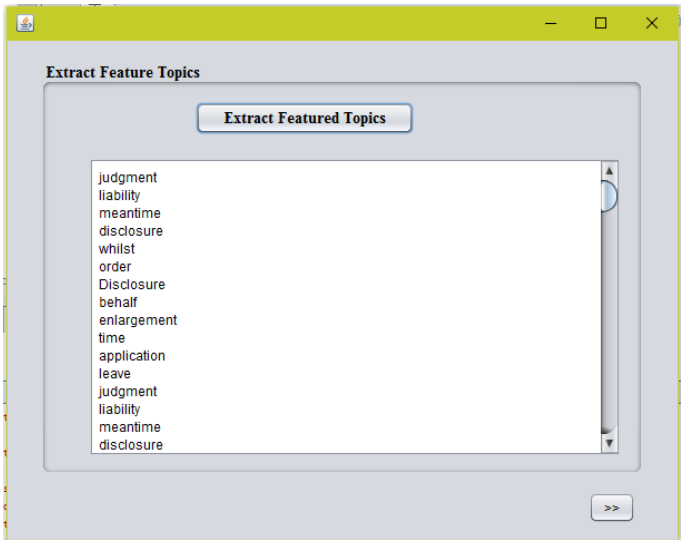


**Figure 4.** Extraction Process.

Figure 4 shows the words present in the document after eliminating the stop words.

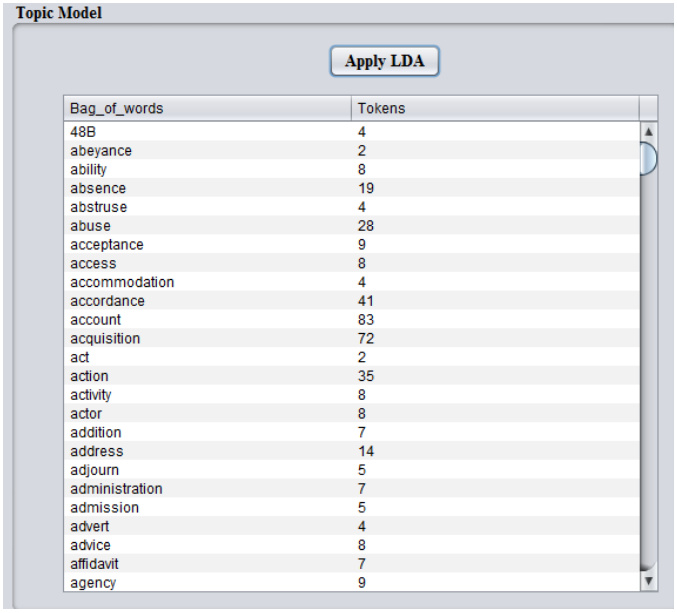


**Figure 5.** Applying LDA.

Figure 5 shows the number of times a word has occurred in the document.

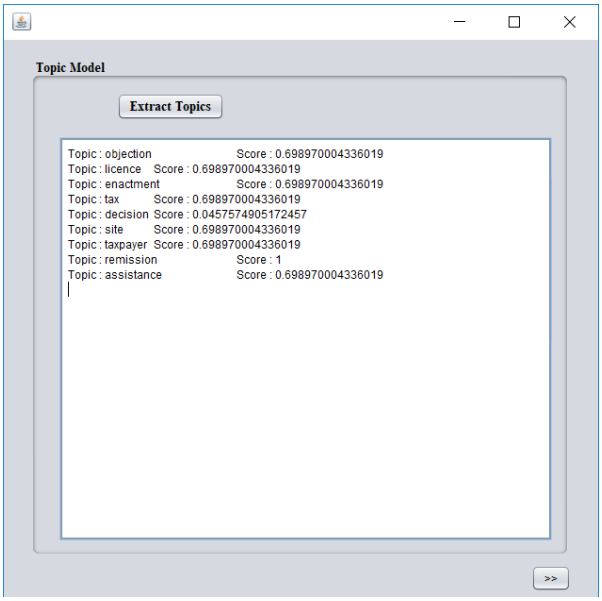


**Figure 6.** Finding Score Calculation.

Figure 6 shows the extraction of topic and score.

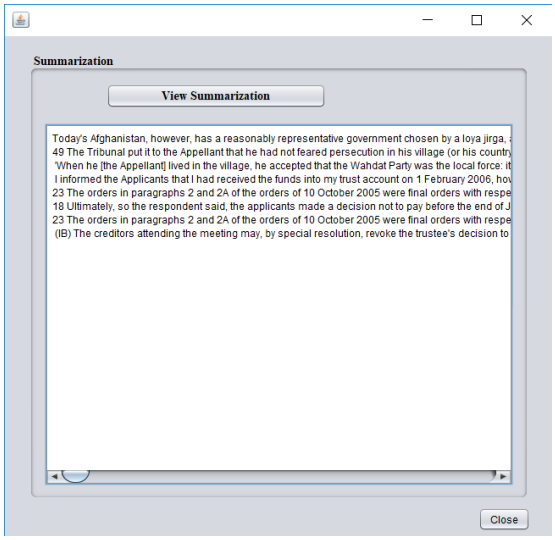


**Figure 7.** Summarization view.

Figure 7 shows the summarization result.

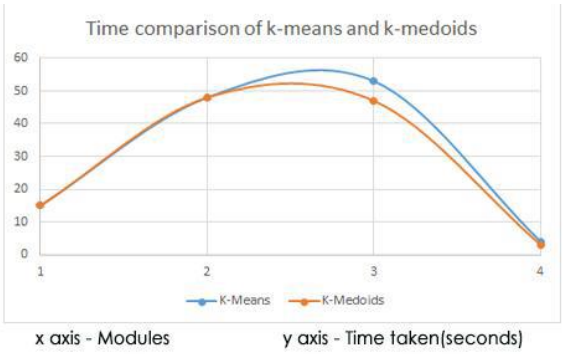


**Figure 8.** Performance Analysis Graph.

Diagram 8 shows the examination between *k*-means and *k*-medoids clustering algorithms, which are utilized during multi-report summarization. The blue line shows the *k*-means algorithm, while Redline demonstrates the *k*-medoids algorithm strategy. X pivot in the diagram signifies the summarization modules, and *y* hub means the time taken to finish the cycle. As spoken to in the diagram, by utilizing the *k*-medoids algorithm, the time utilization in clustering is less when contrasted with the *k*-means algorithm.

## 5. Conclusion

In this paper, an extractive-based big data summarization method has been proposed both for single or multiple archives. In this summarization, some significant sentences are separated from the first document (s). We have contrasted the outcomes and k-means and *k*-medoids algorithm and estimated the run-time unpredictability that shows the proposed method's presentation is improved. As per the aftereffect of the proposed procedure, we can presume that it decreases the excess and gives better summarization. Because of the LDA Model, the topic-conveyance of archives and the term-appropriation of topics are gotten. Experiments show that with the strategy, the summarization has a special exhibition. The rouge esteems improved as per the sentences generative cycle, computing topic-significance, and sentences' topic-conveyance. In the Future, we improve the exactness for better outcomes.

## References

[1] K. T. Belerao and S. B. Chaudhari, Summarization using MapReduce framework based big data and hybrid algorithm (HMM and DBSCAN). 2017 IEEE International Conference on Power, Control, Signals, and Instrumentation Engineering (ICPCSI). doi:10.1109/icpcsi.2017.8392320 (2017).

[2] Y. Chen, H. Chen, A. Gorkhali, Y. Lu, Y. Ma and L. Li, Big data analytics and big data science: a survey, Journal of Management Analytics 3(1) (2016), 1-42. doi:10.1080/23270012.2016.1141332.

[3] J. Bian, Z. Jiang and Q. Chen, Research on Multi-document Summarization Based on LDA Topic Model. 2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics, doi:10.1109/ihmsc.2014.130 (2014).

[4] T. Hansmann and P. Niemeyer, Big Data-Characterizing an Emerging Research Field Using Topic Models, 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), doi:10.1109/wi-iat.2014.15 (2014).

[5] R. K. Lomotey and R. Deters, Towards Knowledge Discovery in Big Data, 2014 IEEE 8th International Symposium on Service-Oriented System Engineering, doi:10.1109/sose.2014.25 (2014).

[6] V. N. Gudivada, D. Rao and V. V. Raghavan, Big Data Driven Natural Language Processing Research and Applications, Big Data Analytics, 203-238. doi:10.1016/b978-0-444-63492-4.00009-5 (2015).

[7] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, Data mining with big data, IEEE Transactions on Knowledge and Data Engineering, 26(1) (2014), 97-107. doi:10.1109/tkde.2013.109

[8] J. Chen and F. You, Text Summarization Generation Based on Semantic Similarity, 2020 International Conference on Intelligent Transportation, Big Data and Smart City (ICITBS), doi:10.1109/icitbs49701.2020.00210 (2020).

[9] P. Sethi, S. Sonawane, S. Khanwalker and R. B. Keskar, Automatic text summarization of news articles, 2017 International Conference on Big Data, IoT and Data Science (BID). doi:10.1109/bid.2017.8336568 (2017).