Turkish Journal of Computer and Mathematics Education Vol.12 No.9 (2021), 2385– 2391
Research Article

Comparative Analysis of Tamil and English News Text Summarization Using Text Rank Algorithm

[a]Sarika M, [a]Dr.Rajeswari K C, and [a]Lavanya A P

[a]Department of Computer Science and Engineering, Sona College of Technology, Salem, Tamil Nadu, India -636005

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: The exponential growth of newsgroups has made it more difficult to gain accurate access to a large amount of data. To deal with the massive amounts of data, efficient and effective methods are needed. One such method is text summarization, which presents data in a condensed format. It would be beneficial for readers to be able to get a wide variety of news in a short amount of time if the news is simplified.In this article, we use the English Newsgroup datasets and the Tamil Newsgroup datasets to automate News Summaries using the text rank algorithm. The proposed work was created using a changed Text Rank algorithm based on the principle of word frequency. The suggested approach creates vectors of words as nodes and similarities between two words as the edge between them, which is the webbing between them. Term frequency assigns various weights to different terms in a sentence, while standard cosine similarity regards both of them similarly. The vector is rendered sparse and divided into clusters based on the premise that sentences inside a cluster are identical and sentences from various clusters reflect their dissimilarity. The performance assessment of the proposed summarization strategy in two types of Newsgroup datasets demonstrates its usefulness in terms of the accuracy parameter.

Keywords: Summarization, News Group dataset, Term frequency, Text Rank algorithm, Vectors.

1. Introduction

Text mining is identified with text examination and is subsequently known as text information mining. Example and pattern development, for example, measurable example learning, is a typical device for giving top-notch information. Text mining is the way toward organizing text data (commonly parsing, with the consideration of some determined etymological highlights and the prohibition of others, and extreme addition into a data set), inferring patterns inside the organized information, lastly breaking down and deciphering the material. In-text mining, great typically alludes to a blend of importance, oddity, and interest. Text mining tasks incorporate information grouping, text bunching, idea/element extraction, feeling examination, data rundown, and substance connection displaying. The fundamental objective is to utilize common language handling (NLP) and factual strategies to interpret text into information for examination. A typical application is to break down a progression of common language archives and afterward either model the record set for factual characterization purposes or populate a data set or search file with the data got. Text examination is the way toward exploring a lot of composed information to give novel thoughts and transform unstructured content into requested information for additional investigation. Text mining uncovers realities, connections, and cases that would some way or another go undetected in an ocean of literary large information. This data is extricated and converted into organized information, which is then utilized for examination, representation (using HTML tables, mind guides, and graphs), joining with organized information in data sets or stockrooms, and further refining utilizing AI (ML) structures. Text mining has gotten more doable for information researchers and different clients as large information frameworks and profound learning calculations that can deal with huge assortments of unstructured information have progressed. Information mining and examination can help organizations in acquiring possibly important industry bits of knowledge from tweets, client messages, call focus logs, verbatim overview results, web-based media posts, and clinical records. As a feature of their marking, dissemination, and client relations exercises, organizations are slowly coordinating content mining innovations into talk bots and menial helpers that offer programmed reactions to clients. Text mining investigation can help organizations in acquiring conceivably significant industry experiences from tweets, client correspondences, call focus logs, verbatim study results, online media posts, clinical records, and other content-based information sources. Organizations as a component of their showcasing, conveyance, and client support exercises, organizations are continuously coordinating content mining innovation into AI talk bots and virtual specialists that give robotized reactions to clients. Text mining is like information mining, however, it centers around text instead of more organized information. Notwithstanding, one of the initial phases in the content mining measure is to arrange and structure the information with the goal that it very well may be utilized for both subjective and quantitative examinations. Regular language handling (NLP) innovation, which utilizes computational phonetics standards to parse and decipher informational collections, is generally utilized in this limit. Text is characterized, grouped, and labeled; informational collections are summed up; scientific categorizations are made; and data about articles, for example, word frequencies and information element connections are recovered.

2.Related work

This method was used by (G. Brown, 2012). We extract feature significance indices from a given objective function rather than attempting to describe them. The conditional probability of the class labels provided the features is a well-accepted mathematical theory that we use as our goal. As a result, we will gain a better understanding of the function selection dilemma and accomplish the goal outlined above: renovating a large number of hand-crafted heuristics into a theoretical context. In this part, we compare and contrast some of the parameters used in the literature. (Ribeiro, 2016) proposed as a solution to the "trusting a forecast" problem, giving explanations for particular predictions, and choosing several such predictions (and explanations) as a solution to the "trusting the model" problem. The following is a list of our major contributions. LIME is an algorithm that can accurately describe the predictions of any classier or regressed by approximating it locally using a model that can be understood. (Zuradaet 2015) looked at neural networks that were configured for classification and were discriminatively educated. The input data is assumed to have nonnegative values. In fact, this criterion is often met. The suggested method necessitates the specification of the network's architecture as well as five parameters that govern their generalization and sigmoid steepness. The number of mystery neurons was chosen to accomplish high characterization accuracy while keeping the organization restricted. The technique recommended by (J. H. Lau, 2011) in this paper is to deliver a theme mark up-and-comer assortment by (1) sourcing subject name applicants from Wikipedia by questioning with the top-N point terms; (2) choosing the highest level record titles; and (3) post-preparing the archive titles to extricate sub-strings. Our commitments to this work include: (1) the advancement of a novel subject name evaluation measure and dataset; (2) the proposition of a strategy for both distinguishing and positioning theme mark up-and-comers; and (3) great in-and get area discoveries through four diverse paper assortments and related point models, showing our technique's capacity to naturally name subjects. In a text retrieval mission, (Aletras 2014) compared various subject representations. We want to know how different subject representation modalities affect identifying appropriate documentation for a question, as well as how difficult it is to view the same topics using different representation modalities. The aim of the challenge was to find as many documents as possible that were applicable to a series of queries. Subjects were asked to complete the retrieval process in two steps. They were given a questionnaire and a list of LDA topics represented by a particular modality (keywords, textual mark, or image).

3. Existing Methodologies

Clustering is defined as devising a classification scheme for grouping objects into many classes, each of which is represented by a series of numerical measurements, such that objects within classes are identical in certain ways but different from those in other classes. It is necessary to decide the number of classes and the features of each class. Clustering is a form of automated learning that divides a collection of objects into subsets or clusters. The aim is to make clusters that are internally consistent but significantly different from one another. The random forest is one of the existing method. The random forest can be defined in figure 1.

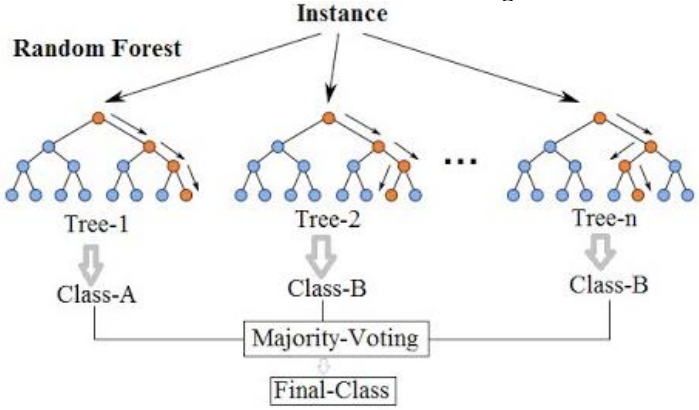

Figure 1: Random Forest Algorithm.

4. Proposed methodology

The proposed exploration would investigate a framework for making synopses from Tamil and English news posts. The essential point is to build up an exact and productive strategy that utilizes the content position

calculation, to sum up given content records as a significant concentration of the first content archive. This examination centers around making rundowns utilizing a book rank calculation, which depicts an assortment of terms that are measurably critical for a report. In a multi-dimensional space, each sentence in content is treated as a vector. The main sentences are those that are nearest in position worth to the position esteem. Three measurements choose the significance of a sentence: the position esteem, the positional worth, and the primary sentence cover. The TR calculation is utilized to quantify the score for each sentence and limit excess between them. At last, the sentences are evaluated, and the synopsis sentences with the most elevated scores are picked. Furthermore, grow the strategy to consolidate the Text Rank calculation into English news datasets to survey the framework's presentation regarding precision boundaries.

4.1 Text Rank Algorithm

Text rank is an unsupervised extractive information summarization methodology. Each sentence in the document is assigned a particular rank or ranking using the text rank algorithm. The value of a sentence in the document is indicated by its rank. This suggests that the higher the sentence's rank, the more relevant the sentence is. Sentences with a rank greater than or equal to a certain threshold are considered for summary generation. Text mining is a technique for extracting useful information, knowledge, or patterns from text documents from a variety of sources. It contains following steps as follows

- Step 1: Choosing the document's scope
- Step 2: Tokenization
- Step 3: Token Normalization
- Step 4: Remove stop words
- Step 5: Phrase stemming
- Step 6: Get rid of all unique characters
- Step 7: The similarity of each sentence to the previous sentence is determined
- Step 8: Use the Text rank algorithm to calculate the score for each sentence.
- Step 9: Select sentences with the highest scores to be included in the summary.

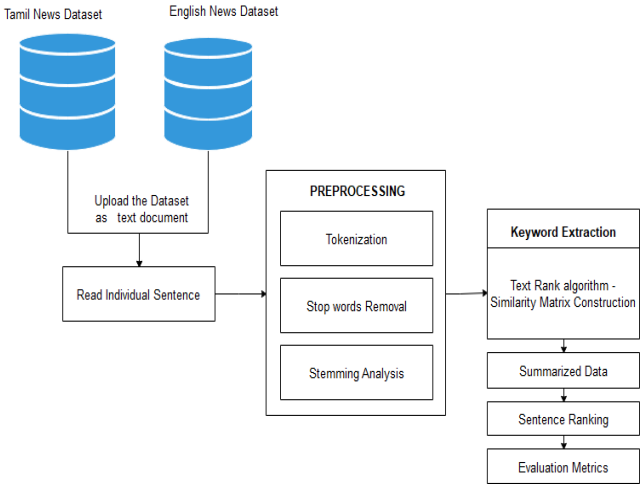

Figure 2: Proposed Framework

The system's overall configuration was depicted in the diagram above. The datasets as English and Tamil news datasets can be uploaded as CSV files by the administrator. Then interference terms are then removed and the stemming term is analyzed using data pre - processing procedures. Finally, as functions, extract the keywords. A similarity matrix will be computed using this formula, and the results will be summarized with greater precision.

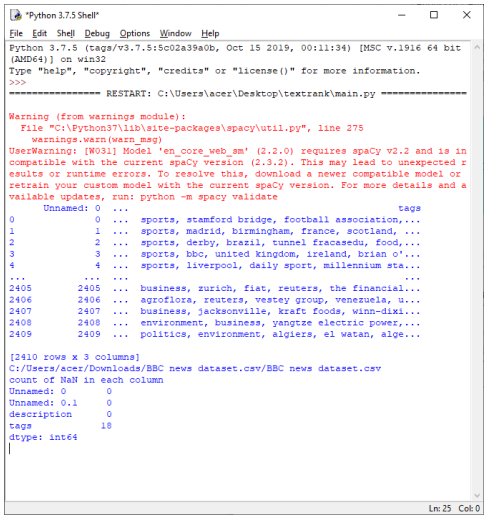

Figure 3: Data preprocessing for English News

Since regional language summaries is a developing phenomenon, there are only a small number of libraries available, and the open-Tamil library is used to summarize Tamil news. Raw Tamil news will be fed into this system as an input. Data will be tokenized into sentences and words in this process, and character constructs such as punctuation, time, date numbers, and so on will be detected. After that, stop words such as "gy (pala)", "xU (oru)", "vd; W (enru)", and so on can be omitted by measuring the frequency of the tokens; for this, a list of punctuation marks in Tamil will be used.

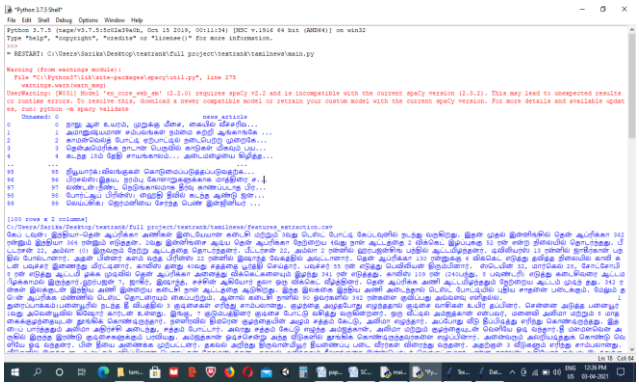

Figure 4: Data preprocessing for Tamil News

4.2 Open – Tamil Library

The open-Tamil library, for example, is designed to develop high-level applications in Tamil. The Open-Tamil library is a free Python 2 (and Python 3K) package that can be used to process Tamil text. These Open-Tamil lessons can be applied to other Indian languages on the internet. Instead of thinking about encodings and code-points, users can manipulate Tamil text at the letter level, resulting in a stronger separation of architecture and detail. Independent of encodings, the API allows for streaming algorithms on canonical Tamil results.

5. Experimental result

The text rank algorithm is applied to Newsgroup datasets in this article. Precision, Recall, F1 Score, and Accuracy Parameter can all be used to predict the system's results. The main assessment metrics of co-selection measures are precision, recall, and F-score.

Precision (P) is calculated by dividing the number of sentences in both the candidate and comparison summaries by the number of sentences in the candidate summary.

Precision =TP/(TP+FP)

Recall (R) The number of matched sentences in both the nominee and reference summaries is divided by the number of sentences in the reference summary to calculate recall (R).

Recall =TP/(TP+FN)

F-score The Precision and recall are combined in the F-score. A harmonic average of accuracy and recall is what the F-score is.

F measure = 2* (Precision*Recall)/(Precision+Recall)

1.Weighted average of precision, recall, F1-score for Tamil news summarization:

Precision	0.50
Recall	0.67
F1-score	0.80

2.Subjective Measure for Tamil news summarization:

Subject	Time is taken to read all the news (before summarization) (in mins)	Time is taken to read the news(after summarization) (in mins)
S1	29	10
S2	38	14
S3	35	12
S4	40	19
S5	47	21
Mean	37.8	15.2

The time it takes to read the summary news and the time it takes to read all of the news (before summarization) is 15.2:37.8. 22.6 is the disparity between the two means. The figure is 59.78 percent as calculated. As a result of this finding, it is clear that reading the summary news instead of reading all of the news will save 59.78 percent of time.

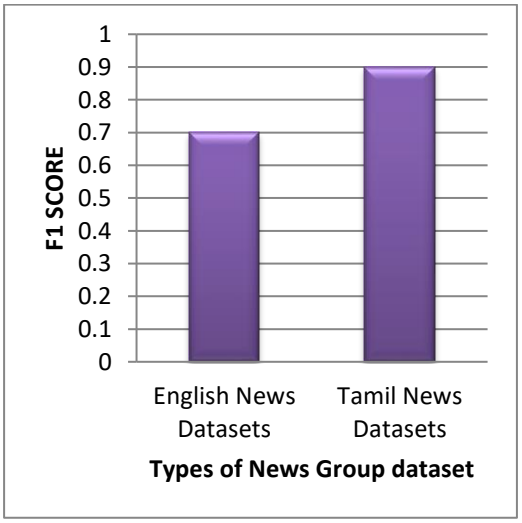

Figure 5: F1 Score chart

Also, the Accuracy parameter was used to measure the efficiency of text-analysis programs. The fraction of the total number of perfect predictions to the total number of test data is used to calculate accuracy (ACC). 1 – ERR is another way to do it. The best possible accuracy is 1.0, although the lowest possible accuracy is 0.0.

$$ACC= (TP+TN)/(TP+TN+FN+FP) \times 100$$

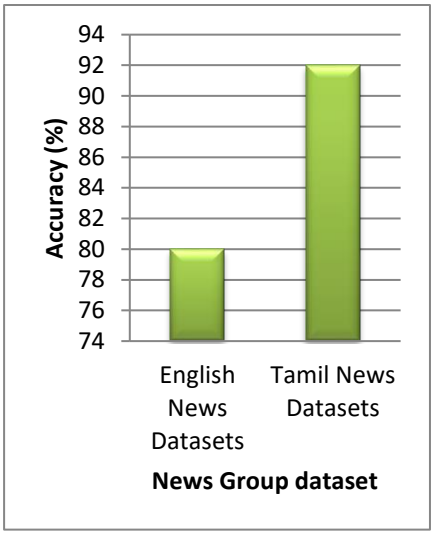

Figure 6: Accuracy chart

From the above figures, the highest performance is obtained (accuracy and F1 score) in the Tamil newsgroup dataset summarization.

6. Conclusion

Text summarization is essential in both the commercial and academic communities. Automatic text summarization has become possible to accomplish, particularly using Python, thanks to the availability of numerous libraries. In this study, we found that our proposed scheme, which uses Text Rank as a similarity metric, performed worse than current similarity vectors. As compared to the other similarity measures measured for two newsgroup datasets, semantic folding can be considered an option for similarity measure in Text Rank since the discrepancy between the measures is so small. In terms of Accuracy and F1 score parameters, experimental findings showed that the Text Rank algorithm worked well in Tamil News Group datasets.

7. Future works

The proposed structure would be expanded in the future to include another Indian regional language. The Tamil news summarization in this paper is achieved with the open – Tamil python library; however, there are some limitations in the preprocessing and function extraction processes that will be addressed in future work in order to have even more consistency in summarization.

References

1. Aletras, Nikolaos, et al. (2014) Representing topics labels for exploring digital libraries. Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. IEEE Press.
2. Agarwal B, Mittal N. (2014) Text classification using machine learning methods-a survey. In Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012 (pp. 701-709). Springer, New Delhi
3. Arunkarthikeyan, K. and Balamurugan, K., 2020, July. Performance improvement of Cryo treated insert on turning studies of AISI 1018 steel using Multi objective optimization. In 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE) (pp. 1-4). IEEE.
4. Aroulanandam, V.V., Latchoumi, T.P., Bhavya, B., Sultana, S.S. (2019). Object detection in convolution neural networks using iterative refinements. Revue d'Intelligence Artificielle, Vol. 33, No. 5, pp. 367-372. https://doi.org/10.18280/ria.330506
5. Brown, Gavin, et al. (2012) Conditional likelihood maximization: a unifying framework for information-theoretic feature selection. Journal of machine learning research;13: 27-66.

6. Bhargava R, Sharma Y, Sharma G. (2016) Aussi: Abstractive text summarization using sentiment infusion. Procedia Computer Science; 89: 404-411.
7. Bhasha, A.C. and Balamurugan, K., 2020, July. Multi-objective optimization of high-speed end milling on Al6061/3% RHA/6% TiC reinforced hybrid composite using Taguchi coupled GRA. In 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE) (pp. 1-6). IEEE.
8. Chien, Jen-Tzung. (2016) Hierarchical theme and topic modelling. IEEE transactions on neural networks and learning systems; 27(3):565-578.
9. Chinnamahammad Bhasha, A., Balamurugan, K. Fabrication and property evaluation of Al 6061 + x% (RHA + TiC) hybrid metal matrix composite. SN Appl. Sci. **1,** 977 (2019). https://doi.org/10.1007/s42452-019-1016-0
10. Chorowski Jan, Jacek M, Zurada (2015) Learning understandable neural networks with nonnegative weight constraints. IEEE transactions on neural networks and learning systems;26(1): 62-69.
11. Deepthi, T. and Balamurugan, K., 2019. Effect of Yttrium (20%) doping on mechanical properties of rare earth nano lanthanum phosphate (LaPO4) synthesized by aqueous sol-gel process. Ceramics International, 45(15), pp.18229-18235.
12. Garikipati P., Balamurugan K. (2021) Abrasive Water Jet Machining Studies on $AlSi_7$+63%SiC Hybrid Composite. In: Arockiarajan A., Duraiselvam M., Raju R. (eds) Advances in Industrial Automation and Smart Manufacturing. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-15-4739-3_66
13. Gowthaman, S., Balamurugan, K., Kumar, P.M., Ali, S.A., Kumar, K.M. and Gopal, N.V.R., 2018. Electrical discharge machining studies on monel-super alloy. Procedia Manufacturing, 20, pp.386-391.
14. Kummamuru, Krishna, et al. (2004) A hierarchical monothetic document clustering algorithm for summarization and browsing search results. Proceedings of the 13th international conference on World Wide Web. ACM.
15. Kurian N, Asokan S. (2015) Summarizing user opinions: A method for labeled data-scarce product domains. Procedia Computer Science;46: 93-100.
16. Lau, Jey Han, et al. (2011) Automatic labelling of topic models. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Human Language Technologies (1). Association for Computational Linguistics.
17. Ranjeeth, S., Latchoumi, T.P., Sivaram, M., Jayanthiladevi, A. and Kumar, T.S., 2019, December. Predicting Student Performance with ANNQ3H: A Case Study in Secondary Education. In 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 603-607). IEEE.
18. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. (2016) Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM.
19. Scaiella, Ugo, et al. (2012) Topical clustering of search results. Proceedings of the fifth ACM international conference on Web search and data mining; ACM.
20. Tseng, Yuen-Hsien. (2010) Generic title labeling for clustered documents. Expert Systems with Applications; 37(3): 2247-2254.
21. Xie, Pengtao, Eric P. Xing. (2013) Integrating document clustering and topic modeling. arXiv preprint arXiv;1309.6874.
22. Ghosh, D. (2020) A Sentiment-Based Hotel Review Summarization. In Emerging Technology in Modelling and Graphics Springer, Singapore; 39-44.
23. Yan Z, Xing M, Zhang, D, Ma B. (2015) EXPRESS: An extended PageRank method for product feature extraction from online consumer reviews. Information & Management;52(7): 850-858.