Grenze International Journal of Engineering and Technology, Jan Issue

News Classification and Summarization using Random Forest and TextRank Algorithm

Jaydeep Patel[1], Hardi Patel[2], Mitali Patil[3] and Mr. Harsh Namdev Bhor[4]
[1-3]UG Student, K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, India
Email: jaydeep.patel@somaiya.edu, hardi.p@somaiya.edu, mitali.patil@somaiya.edu
[4]Asst. Professor, K. J. Somaiya Institute of Engineering and Information Technology, Sion, Mumbai, India
Email: hbhor@somaiya.edu

***Abstract*—We read numerous news content every day, yet a single news subject may not include all of the important information. We can obtain useful information from the comprehensive articles, which will save us a lot of time and energy. Text summarization is a method of compressing a lengthy text into a comprehensible summary while preserving the information content and its overall meaning. This system will help us enhance the readability of a particular news article without missing any important points. To understand the sentiment of the news items, Sentiment Analysis is conducted on the summary. Text classification is used to organize, structure, and categorize different types of news articles.**

***Index Terms*— Text Summarization, Text Classification, Sentimental Analysis, Random Forest, TextRank algorithm, Exploratory Data Analysis, Data pre-processing, Natural Language Processing, Machine Learning.**

I. INTRODUCTION

Giving computers the ability to think has always been a pipe dream for humans, but with the advancement of Machine Learning in recent decades, giving computers a "brain" is no longer a pipe dream. Yes, machines may now be trained to learn and apply their knowledge to do formerly human-only activities. Machine Learning applications include creating a summary of a given piece of text and converting text to speech. As a consequence, it is evident that this study applies a machine learning-based technique to produce a clear and brief summary of a block of content. Consequently, it makes a significant contribution to and has the ability to contribute to a wide range of disciplines. It also aids a number of domains and may be linked to or implemented in a number of other systems. Additional aim is to enable visually impaired persons to acquire news much more efficiently without the assistance of others. Text Reduction, also known as Automatic Text Summarization, is an intriguing and rapidly emerging field of natural language processing.

Grenze ID: 01.GIJET.9.1.110

II. LITERATURE REVIEW

TABLE I. OVERVIEW OF THE REVIEWED SOURCES

Sr. No	Title	Authors	Description
1.	Text Summarization: An Overview	Mr. S.A. Babar	This paper primarily deals with 2 key issues: Finding the appropriate document Amount of such paper accessible for that precise information. Based on the above-mentioned issues, this paper aims at inventing a system for automatically summarizing the text which is crucial for addressing the challenges outlined in the research [1]
2.	Extractive Text Summarization Using Sentence Ranking	J.N. Madhuri, Ganesh Kumar.R	This research paper demonstrates a novel approach to summarize text using the Extractive based method. The key idea behind the approach is to assign weights to the sentences which is then used to rank those sentences. Subsequently, high rated sentences are extracted from the input to generate a high-quality summary. An evaluation of the proposed system is done by comparing the results obtained by the system w.r.t Human generated summary and M.S. Word generated summary [2]
3.	Automatic Text categorization and summarization using rule reduction	C. Lakshmi Devasena, M. Hemalatha	To assess the structure of input text, this study employs an automatic text classification and summarization technique. The authors successfully created a text analyzer that used a rule reduction approach to extract the structure of the input text in three different stages: token creation, feature identification and categorization, and summarization [3]
4.	An Overview on Extractive Text Summarization	Shohreh Rad Rahimi, Ali Toofanzadeh Mozhdehi, Mohamad Abdolahi	The issue of text mining and its link with text summarization is explored first in this paper. Then, an exploration of some of the summary systems and their critical characteristics for extracting dominating phrases was performed. The major stages of the summarizing process were identified, and the most significant extraction criteria were provided. Finally, the most basic recommended assessment techniques are taken into account [4]
5.	Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence	P.Krishnaveni, Dr.S. R. Balasundaram	The authors developed a heading wise summarizer, an automated extractive text summarizer, in their suggested technique. It produced a statistical and linguistic feature-based extractive single document heading-wise summary of the source material. The experiment results reveal that heading wise summarizer outperforms the main summarizer, free summarizer, auto summarizer, and MS. word summarizer in terms of precision, recall, and f-measure [5]
6.	Extractive Text Summarization- An effective approach to extract information from text	Asha Rani Mishra, V.K Panchal, Pawan Kumar	This study focuses on primary strategies for extracting relevant information from a given text using topic modelling, key word extraction, and summary construction. The LSI and NMF methods are deployed for topic modelling, the weighted TF-IDF method is used for key word extraction, and the LSA and Text Rank methods are employed to construct text summaries [6].
7.	Two-level Text Summarization from Online News Sources with Sentiment Analysis	Tarun B. Mirani, Sreela Sasi	An extractive-based technique is applied in this study to construct a two-level summary from internet news items. Politics, sports, health, science, and movie reviews are among the themes covered by Fox News in the United States, the NZ Herald in New Zealand, the Hindustan Times in India, and the BBC in the United Kingdom, among others. The very first summary creates a summary for each of these articles. Sentiment Analysis is used to the first-level summary in order to comprehend the variance in related news stories from various media sources. The second level summary creates the summary of the combined first-level summaries of multiple linked articles on a topic. The ROUGE measure is used to assess summarization performance [7].
8.	A general decision layer text classification fusion model	Xiao-Dan Zhang	This study proposes a novel decision classification fusion model and method. Texts are fed into the local fusion centre, and their findings are fed into the decision layer. As feature fusion algorithms, we use KNN, SVM, and BP Net, and the D-S Theory method as the decision layer. This approach uses the identical training and test sets for KNN, SVM, BP Net, and the fusion model to demonstrate the model's performance [8].

III. METHODOLOGY

For the inputted news article, the system generates an extractive-based summary. The article is first pre-processed, and then a summary is generated. The generated summary is then passed to the sentiment analyzer to determine its polarity. Concurrently, the news article is run through a text classification algorithm to determine which category it belongs to. Finally, the summary is available in audio format for the benefit of the visually impaired. The major stages of the system are described below.

The proposed system contains two major phases:

1. News classification
2. News summarization.

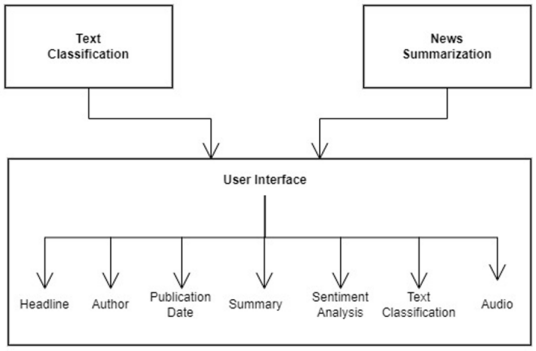

Figure 1.Displays how the various phases of the system interact with each other while creating the final summary

A. News Summarization

Concatenating all of the material in the articles would be the first stage, followed by separating the text into individual phrases. The next stage is to find vector representations (word embeddings) for each phrase. The computation of sentence rank is accomplished by transforming the similarity matrix to a graph with sentence fragments as vertices and similarity scores as sides. Finally, the final summary is made up of a predetermined amount of top-ranked sentences.

B. News Classification

Classification is a technique for categorizing natural language texts based on their content. Topic Spotting is another term for text categorization. By using several machine learning algorithms, this experimental text classification solution seeks to appropriately categorize the inputted news item from a broad variety of categories such as business, technology, politics, sports, and entertainment. [10]

IV. IMPLEMENTATION

A. News Categorization

Exploratory Data Analysis

In this research we have used exploratory data analysis (EDA) as a way of evaluating datasets to summarise essential properties utilising visual approaches.

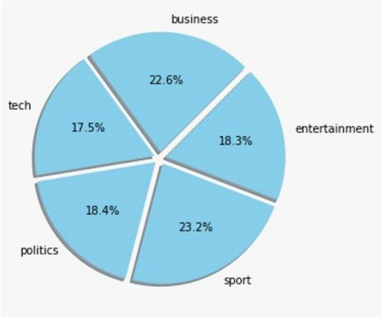

Figure 2. Percentage of articles in different categories

Figure 3. Word cloud for business category

The graph seen above in Fig. 2. is a pie chart that displays the percentage of articles in the following five categories: business, technical, entertainment, politics, and sports. The chat indicates that the most articles in the dataset are in the sector of sport, followed by business and politics. The technology category has the lowest percentage of articles.
The term "Word Cloud" refers to a data visualization approach for visualizing text data in which the size of each word represents its frequency or relevance. A word cloud for the category "business" is shown in Fig. 3.

Data cleaning and pre-processing
The process of turning raw data into a comprehensible format is known as data preparation. This is an important stage in data mining since we cannot deal with raw data. Before utilising machine learning or data mining methods, the data quality should be evaluated.
The following step was performed in this approach to clean and pre-process the data:

1. Remove all tags
2. Remove special character
3. Convert everything to lowercase
4. Stop-word removal
5. Lemmatization

Model Building
1. Train/Test data split
For this experimental research, we have utilized a train test split of the sklearn.model selection module. This data was divided into labels and features by the approach. The parameters test size=0.3 and shuffle is toggled to true.
2. Create and Fit Model
Various classification algorithms have been used for this experiment
2.1. Logistic Regression (Classification Technique 1)
Logistic Regression is a classification technique notable for its exponential and log-linear functions. It works with discrete values and transfers the function of any real value into 0 and 1. The hypothesis for sentiment analysis demonstrates that reviews can be good or negative.
2.2. Random Forest (Classification Technique 2)
We utilized RandomForestClassifier from the sklearn.ensemble module. Starting with parameters like n estimators, criterion, and random states, we constructed a Gaussian Classifier. The trees in this approach are counted using "n estimators." That parameter has been kept at its default value of 100. A 'criterion' parameter is a metric that is used to assess the quality of a split. This option has been set to entropy.
2.3. Multinomial Naive Bayes (Classification Technique 3)
MultinomialNB was imported from the sklearn.naive bayes module of sklearn to create the Multinomial Naive Bayes model. We did not use any of the numerous accessible parameters, such as alpha, fit prior, or class prior, in this experiment. The model is subsequently trained using fit() and the training set of data. Then, we projected and produced output using the testing data and the predict function.
2.4. Support Vector Classifier (Classification Technique 4)
SupportVectorClassifier was imported from the sklearn.svm package. First, we created an SVC model. The model was then trained and forecasted using the.fit() and.predict() functions, respectively.
2.5. Decision Tree Classifier (Classification Technique 5)
We utilized DecisionTreeClassifier from sklearn.tree for this project. The model was initialized, trained and predicted like the other models mentioned above.
2.6. K Nearest (Classification Technique 6)
KNeighborsClassifier was imported from sklearn.neighbors. We started by building a classifier with parameters like n_neighbors, metric, and p. n_neighbors used to specify the number of neighbors in KNN. The parameter metric is used at its default value while p which is the power parameter is set to 4.
2.7. Gaussian Naive Bayes (Classification Technique 7)
This classification technique is computed like the Multinomial Naive Bayes algorithm.

Model Evaluation
Model is evaluated on the basis of a four-parameters:
1. Accuracy
The percentage of correct predictions for the test data is known as accuracy. It is simple to compute by simply dividing the number of right guesses by the total number of forecasts.
2. Precision

This is the fraction of all observations that were projected to be positive and are really positive.

3. Recall

This is the proportion of observations anticipated to be in the positive category that are really in the positive category. It implicitly indicates the model's ability to identify an observation that belongs to the positive class at random.

4. F1 Score

Precision and recall are inversely related complementing measurements. If we want both, we'll utilize the F1 score to integrate accuracy and recall into a single statistic.

B. News Summarization

a. Exploratory Data Analysis

Deriving insights from raw statistics can be time-consuming, uninteresting, and/or daunting. Exploratory data analysis approaches have been developed to assist in this circumstance.

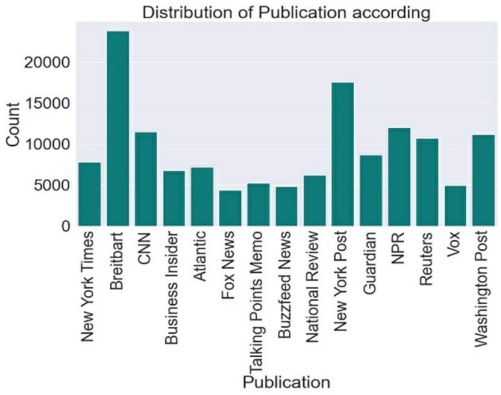

Figure 4. News publication distribution

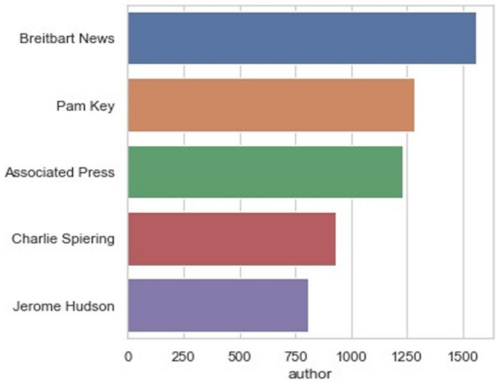

Figure 5. Count of articles published by different authors

b. Data Cleaning and Pre-processing

In this technique, the following steps were taken to clean and pre-process the data:

1. Converting to lowercase
2. Removing the HTML
3. Removing the email ids
4. Removing The URLS
5. Removing the contractions
6. Removing the Trailing and leading whitespace and double spaces
7. Removing punctuations from the article
8. Removing the stopwords [11].
9. Keyword Extraction

c. Summarization

Below figure shows the flow chart of the text rank algorithm,

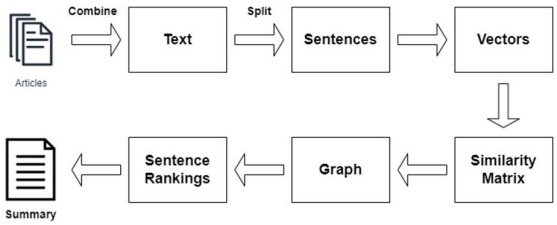

Figure 6. News summarization using text classification

Concatenating all of the material in the articles would be the first stage, followed by separating the text into individual phrases. The similarities between sentence vectors are then calculated and kept in a matrix. The similarity matrix is then graphed to calculate sentence rank, with sentences acting as vertices. The similarity

matrix is then graphed for sentence rank calculation, with sentences acting as vertices and similarity scores acting as edges. Finally, the final summary is made up of a certain number of top-ranked sentences.

V. RESULT

After using the BBC news data set for text classification and applying different algorithms and analyzing them using 4 parameters- Model accuracy, precision, recall, and f1-score we came to the conclusion that Random Forest was the best algorithm with a test accuracy of 97.99% [ref. TABLE-1].
Implemented web app for users using Flask where we can paste link of the news article and obtain authors name, category of the news, summary and sentiment analysis of the text.

TABLE I. DATA FRAME TO COMPARE DIFFERENT MODEL PERFORMANCES

Sr No.	Model	Test Accuracy	Precision	Recall	F1 Score
1	Logistic Regression	97.32	0.9741	0.9711	0.9722
2	Random Forest	97.99	0.9794	0.9793	0.9793
3	Multinomial Naive Bayes	97.99	0.9786	0.9812	0.9793
4	Support Vector Classifier	96.98	0.9696	0.9693	0.9691
5	Decision Tree Classifier	80.87	0.8548	0.8085	0.8170
6	K Nearest Neighbour	76.85	0.8218	0.7624	0.7712
7	Gaussian Naive Bayes	79.53	0.8938	0.7897	0.8100

Figure 7. Index page to submit URL of the article

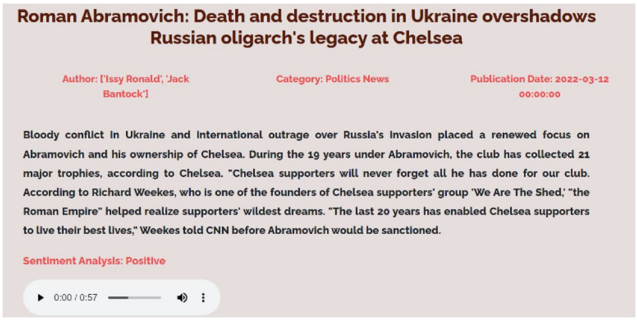

Figure 8. Summary page of the article

VI. CONCLUSIONS

Extraction-based Summarization with News Classification and Sentiment Analysis is presented in this study. The created summary only includes the most significant information from internet news items. News Classification assists us in categorizing such news pieces. Sentiment Analysis presents the perspective of many

news outlets. Finally, the model-summarized material is transformed to voice. Finally, we were able to extract the fundamental synopsis of each news story that was submitted to it. This summary was checked for correctness using the number of keywords retrieved as output and their relevance in the text, and we were able to achieve high efficiency and accuracy, leading the system to the predicted conclusion.

REFERENCES

[1] Babar, Samrat & Tech-Cse, M & Rit,. (2013). Text Summarization:An Overview.

[2] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

[3] C. Lakshmi Devasena and M. Hemalatha, "Automatic Text categorization and summarization using rule reduction," IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012), 2012, pp. 594-598.

[4] S. R. Rahimi, A. T. Mozhdehi and M. Abdolahi, "An overview on extractive text summarization," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017, pp. 0054-0062, doi: 10.1109/KBEI.2017.8324874.

[5] P. Krishnaveni and S. R. Balasundaram, "Automatic text summarization by local scoring and ranking for improving coherence," 2017 International Conference on Computing Methodologies and Communication (ICCMC), 2017, pp. 59-64, doi: 10.1109/ICCMC.2017.8282539.

[6] A. R. Mishra, V. K. Panchal and P. Kumar, "Extractive Text Summarization - An effective approach to extract information from Text," 2019 International Conference on contemporary Computing and Informatics (IC3I), 2019, pp. 252-255, doi: 10.1109/IC3I46837.2019.9055636.

[7] Mirani, Tarun & Sasi, Sreela. (2017). Two-level text summarization from online news sources with sentiment analysis. 19-24. 10.1109/NETACT.2017.8076735.

[8] X. Zhang, "A general decision layer text classification fusion model," 2010 2nd International Conference on Education Technology and Computer, 2010, pp. V5-239-V5-241, doi: 10.1109/ICETC.2010.5529774.

[9] C. Harikumar, “KEYWORD ANALYSIS:USING TEXT ANALYSIS FOR SEARCH ENGINE OPTIMIZATION.” Accessed: Nov. 11, 2020. [Online]. Available: http://www.ijpam.eu.

[10] Z. Li, W. Shang and M. Yan, "News text classification model based on topic model," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016, pp. 1-5, doi: 10.1109/ICIS.2016.7550929.

[11] S. Wankhede, R. Patil, S. Sonawane and P. A. Save, "Data Preprocessing for Efficient Sentimental Analysis," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 723-726, doi: 10.1109/ICICCT.2018.8473277.