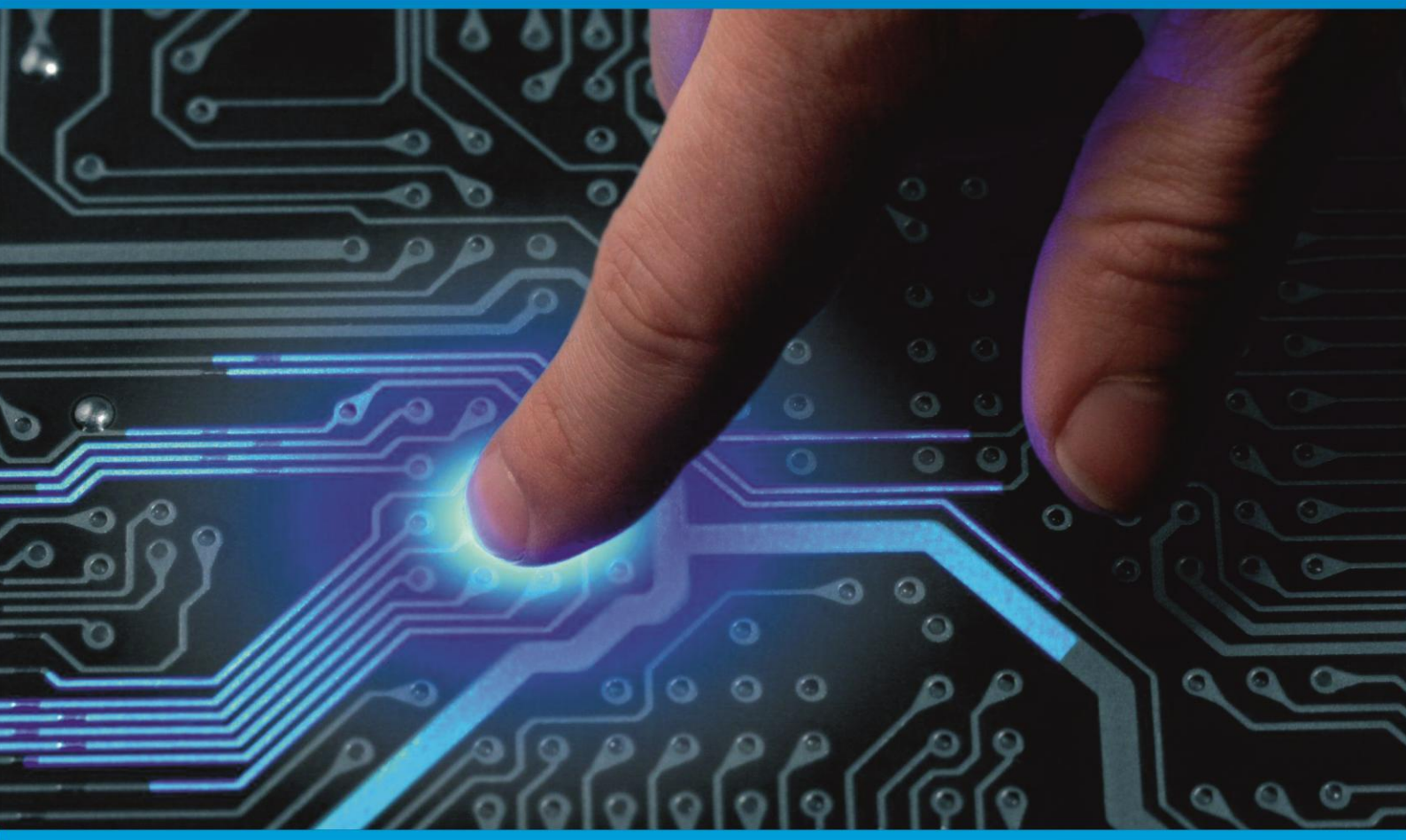
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904167 |

Classification of Summarized News Using SVM

Preetesh Kalshetty, Sanket Pise, Atharva Dhumal, Chetan Patil, Prof. Amruta Kapre

Dept. of Computer Engineering, Zeal College of Engineering and Research Centre, Pune, India

-

ABSTRACT: Text Summarization has consistently been a territory of dynamic interest in the scholarly world. Lately, despite the fact that few procedures have being produced for programmed text outline, proficiency is as yet a worry. Given the expansion in size and number of archives accessible on the web, an effective programmed news summarizer is the need of great importance.

In this paper, we proposed a strategy of text synopsis which centers around the issue of recognizing the main bits of the content and delivering intelligible outlines. Right off the bat, the main component is word highlights, we score each word and separated words that surpassed the preset score as catchphrases and on the grounds that news text is an exceptional sort of text, it contains numerous particular components, like time, spot and characters, so here and there these uncommon news components can be extricated straightforwardly as watchwords.

Second is sentence includes, a direct blend of these highlights shows the significance of each sentence and each component is weighted by Genetic calculation. Finally the summarization is done by machine learning algorithm.

KEYWORDS: Natural Language Toolkit, Support Vector Machine, News Summarization, Natural Language processing

I. INTRODUCTION

Right now, there are huge amounts of printed information accessible, including on the web reports, articles, news, and surveys that contain long strings of text that should be summed up. The importance of text summarization is due to several reasons, including the retrieval of significant information from a long text within a short period, easy and rapid loadingof the most important information.

Currently, automatic text summarization has applications in several areas such as news articles, emails, research papers and online search engines to receive summary ofresults found, So we are proposed news summarization system based on machine learning for giving promising solution with high accuracy. The proposed system consists of news summarization based on document.

1.1 Motivation

The main purpose of the news summarization is to condense the documents or reports into a shorter version and preserve important contents.

Text summarization presents the user a shorter version of text with only vital information and thus helps him to understand the text in shorter amount oftime.

Motive behind proposed work is to achieve higher accuracy over existing work by using machine leering.

The problem with knowledge engineering method is that it requires constant updating of rules for classification which is very difficult. Over the last two decades, the application of Machine learning approach is increased due to various reasons like availability of large amount of data and the necessity of handling them in an efficient way.

1.2 Need

To develop a system that classify soil and suggesting crops with maximum precision and with minimum processing time to help in the law public sector.

II. LITERATURE SURVEY

YAN DU et al.[1] stated that, the automatic text summarization work has become more and more important because the amount of data on the Internet is increasing so fast, and automatic text summarization work can extract useful information and knowledge what user's need that could be easily handled by humans and used for many purposes. Especially in people's daily life, news text is the type of text most people are exposed to. In this study, a new automatic summarization model for news text which based on fuzzy logic rules, multi-feature and Genetic algorithm (GA) is introduced. Firstly, the most important feature is word features, we score each word and extracted words that exceeded the preset score as keywords and because news text is a special kind of text, it contains many specific elements, such as

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904167 |

time, place and characters, so sometimes these special news elements can be extracted directly as keywords. Second is sentence features, a linear combination of these features shows the importance of each sentence and each feature is weighted by Genetic algorithm. At last, we use fuzzy logic system to calculate the final score in order to get automatic summarization. The results of the proposed method was compared with other methods including Msword, System19, System21, System 31, SDS-NNGA, GCD, SOM and Ranking SVM by using ROUGE assessment method on DUC2002 dataset show that proposed method outperforms the aforementioned methods.

G. Bello-Orgaz et al. [2] proposed that Big data has become an important issue for a large number of research areas such as data mining, machine learning, computational intelligence, information fusion, the semantic Web, and social networks. The rise of different big data frameworks such as Apache Hadoop and, more recently, Spark, for massive data processing based on the MapReduce paradigm has allowed for the efficient utilisation of data mining methods and machine learning algorithms in different domains. A number of libraries such as Mahout and SparkMLib have been designed to develop new efficient applications based on machine learning algorithms. The combination of big data technologies and traditional machine learning algorithms has generated new and interesting challenges in other areas as social media and social networks. These new challenges are focused mainly on problems such as data processing, data storage, data representation, and how data can be used for pattern mining, analysing user behaviours, and visualizing and tracking data, among others. In this paper, we present a revision of the new methodologies that is designed to allow for efficient data mining and information fusion from social media and of the new applications and frameworks that are currently appearing under the "umbrella" of the social networks, social media and big data paradigms.

A. S. Nengroo et al. stated that [3] Content summarization is the way toward shortening the source archive into dense structure keeps generally thought regarding the record. The systems of content summarization are abstractive and extractive. The abstractive summarization requires characteristic language preparing devices for outlining the records. The extractive summarization requires factual, etymology and heuristics strategies for positioning the sentences. Numerous strategies have been produced for the summarization of content in different dialects. This paper examine about the strategies for abstractive & extractive text summarization.

A. Sinha et al. [4] proposed that Advertisement identification and filtering in web pages gain significance due to various factors such as accessibility, security, privacy, and obtrusiveness. Current practices in this direction involve maintaining URL-based regular expressions called filter lists. Each URL obtained on a web page is matched against this filter list. While effectual, this procedure lacks scalability as it demands regular continuance of the filter list. To counter these limitations, we devise a machine learning based advertisement detection system using a diverse feature set which can distinguish *advertisement blocks* from *non-advertisement blocks*. The method can act as a base to provide various accessibility-related features like smooth browsing and text summarization for persons with visual impairments, cognitive impairments, and photosensitive epilepsy. The results from a classifier trained on the proposed feature set achieve 98.6% accuracy in identifying advertisements.

H. Saggion et al. [5] proposed that the text Summarization has been an extensively studied problem. Traditional approaches to text summarization rely heavily on feature engineering. In contrast to this, they propose a fully data-driven approach using feedforward neural networks for single document summarization. We train and evaluate the model on standard DUC 2002 dataset which shows results comparable to the state of the art models. The proposed model is scalable and is able to produce the summary of arbitrarily sized documents by breaking the original document into fixed sized parts and then feeding it recursively to the network.

III. PROPOSED METHOD AND ALGORITHM

1. Proposed Methodology:

We proposed the new news summarization based on natural language processing and news classification based on support vector machine model with higher accuracy. We are solve accuracy issue in classification of news articles with accurate stage predictions. The news summarization is done by using natural language toolkit. Almost every Natural Language Processing (NLP) task requires text to be preprocessed before training a model. Deep learning models cannot use raw text directly, so it is up to us researchers to clean the text ourselves. Depending on the nature of the task, the preprocessing methods can be different. The news classification is done by support vector machine.

The System architecture of the proposed model is shown in fig. 1.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904167 |

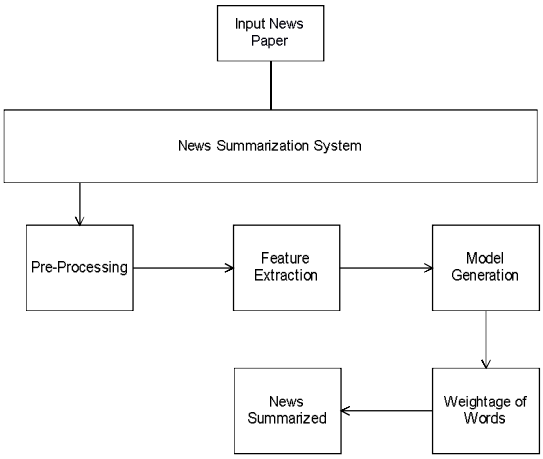


Fig1. Proposed Architecture

The news classification using machine learning algorithm. A sufficient no. of training samples is required. The training samples are collected from fieldwork. The conditions considered while selecting training samples included spatial resolution of the collected data availability of ground reference data, and complexity of the data being considered. The method involves two phases: training phase and testing phase.

2. Algorithms

A. Pre-processing

1. Adding Corpus

This section will load all the email datasets within the program and distribute into training and testing data. This cycle will be accepting the datasets in '*.txt' format for individual email (Ham and Spam). This is to help understand the real-world issues and how might they be tackled.

2. TOKENIZATION

Tokenization is the method where the sentences within an email are broken into individual words (tokens). These tokens are saved into an array and used towards the testing data to identify the event of each word in an email. This will help the algorithms in predicting whether the email ought to be considered as spam or ham.

3. Stop Words Removal

This was used to remove the unnecessary words and characters within each email, and creates a bag of words for the algorithms to compare against. The module 'Check Vectorizer' from Scikit-learn assigns numbers to each word/token while counting and gives its event within an email. The instance is invoked to prohibit the English stopwords, and these are the words such as: A, In, The, Are, As, Is and so on, as they are not exceptionally useful to classify whether the email is spam or not. This instance is then fitted for the program to learn the vocabulary.

4. POS Tagger

Lastly, we can use tag to retrieve the part of speech of each token in a list. Combining all Together We can combine all the preprocessing methods above and create a preprocess function that takes in a *.txt* file and handles all the preprocessing. We print out the tokens, filtered words (after stopword filtering), stemmed words, and POS, one of which is usually passed on to the model or for further processing.

B. Support Vector Machine

This algorithm plots each hub from a dataset within a dimensional plane and through classification strategy the cluster of data is separated by a hyperplane into their particular gatherings shows in equation 1.

$$H = Vx + c \qquad (1)$$

Where c is a constant and V is the vector. The SGD Classifier was loaded from scikit-learn library, which is the linear model with 'Stochastic Gradient Descent (SGD)', also known as the optimized version of SVM. This algorithm gives

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904167 |

more accurate results than SVM (SVC algorithm) itself. Disadvantage of working with SVC algorithm is that it cannot handle a large dataset, whereas SGD gives productivity and other tuning opportunities. The algorithm uses the learning rate to iterate over the sample data to optimize the Linear algorithm and it is signified by the following equation-6 for the default learning rate as 'Optimal' showing in equation2.

$$\frac{1}{\alpha(t_0 + t)} \qquad (2)$$

Where t is the time step which is acquired by multiplying number of iterations with number of samples. The Learning Rate allows implementation of the parameter space during the training time. The α is addresses the regularization term and t0 is a hueristic approach.

IV. CONCLUSION

In this paper, a wide range of strategies needs to investigate in machine Learning and artificial intelligence tailored for news summarization. In this report, we are proposed a new model based on Multi- feature, genetic algorithm for news text summarization. First, extract the most important features (word features and sentence features) and use a linear combination of these features to identify important sentences. In this step, we choose a new text feature extraction method based on the characteristics of news text and summarization using SVM. With this system we provide a user-friendly application that covers aspects like name of that news paper. Using the challenging database in which the news papers are taken. At the front end, i.e. the User Interface, the input will be the news paper by the user. The output will be in the form of summarization of that file.

REFERENCES

[1] H. Saggion and T. Poibeau, "Multi-source, Multilingual Information Extraction and Summarization" in Proc. IEEE Conf. Computer. Vis. Pattern Recognition 2009, pp. 1123–1128.

[2] YAN DU, AND HUA HUO, "News Text Summarization Based on Multi-Feature and Fuzzy Logic" School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China

[3] G. Bello-Orgaz, J. J. Jung, and D. Camacho, ''Social big data: Recent achievements and new challenges,'' *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016.

[4] Ujjwal Rani[1], Karambir Bidhan[2], "Review Paper on Automatic Text Summarization" *CSE* Department, UIET, Kurukshetra University , Kurukshetra , India

[5] A. S. Nengroo and K. S. Kuppusamy, ''Machine learning based heterogenious Web advertisements detection using a diverse feature set,'' Future Gener. Comput. Syst., vol. 89, pp. 68–77, Dec. 2018.

[6] A. Sinha, A. Yadav, and A. Gahlot, ''Extractive text summarization usingneural networks'' Tech. Rep., 2018.

[7] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research [review article]", *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48-57, May 2014.

[8] C.-H. Chen, "Improved TFIDF in big news retrieval: An empirical study", *Pattern Recognit. Lett.*, vol. 93, pp. 113-122, Jul. 2017

[9] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, ''Malicious accounts: Dark of the social networks,'' J. Netw. Comput. Appl., vol. 79, pp. 41–67, Feb. 2017

[10] X. Liu, S. Zhao, A. Liu, N. Xiong, and A. V. Vasilakos, ''Knowledge-aware Proactive Nodes Selection approach for energy management in Internet of Things,'' Future Gener. Comput. Sys.

[11] R. K. Dewang and A. K. Singh, ''State-of-art approaches for review spammer detection: A survey,'' J. Intell. Inf. Syst., vol. 50, no. 2, pp. 234–264, Apr. 2018.

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details