Text Summarizing

Abhiram Prasad
dept. Computer Science
Artificial Intelligence
abhiramprasad@am.students.amrita.edu

Nithin Sylesh
dept. Computer Science
Artificial Intelligence
nithinsylesh@am.students.amrita.edu

Lakshmi G Pillai
dept. Computer Science
Artificial Intelligence
lakshmigpillai@am.students.amrita.edu

Ritika R Prasad
dept. Computer Science
Artificial Intelligence
ritikarprasad@am.students.amrita.edu

Vysakh S Nair
dept. Computer Science
Artificial Intelligence
vysakhsnair@am.students.amrita.edu

***Abstract*—Test summarizing is an effective mechanism for reducing the reading time in this fast phased world with extensive rich overflowing information and data.This research paper encompass an extractive English and Bengali text summarizing tool. A tool that effectively reduces the human effort which produces use-full information by taking a large information or data and discarding unnecessary information.We here have discussed a summarizing implementation based on the application of Machine Learning algorithms based on a set of features extracted from the data-set we have used. And we have used machine learning techniques to aid the computational accuracy of summarizing. Also we have discussed the different application involved in the text summarizing.**

***Index Terms*—Textual Document, formatting, style, styling, insert**

I. Introduction

The amount of large data or textual information that has overflowed in recent years has increased day by day, largely due to news websites, book websites, and a variety of other sources. This can overwhelm the user, requiring extra time and dedication in order for the user to obtain a clear clarity form the textual information.This complicates the process of selecting the appropriate data to optimise. The text summarising approach can successfully handle this problem.
The approach of filtering the most significant details by generating a brief, efficient fluent summary of the textual documents is known as text summarizing. There are two sorts of summarizing techniques: abstractive and extractive. The extractive summarizing approach concatenates significant helpful sentences without comprehending their meaning, but the abstractive summarizing technique creates a shorter meaningful and useful summary. We have used extractive summarizing technique in this paper.Text summarising has a wide range of uses, including media monitoring, news article summarization, search marketing and SEO, internal document management, financial research, and many others.

To generate text summarizing, we used extractive and single document methods. Also various pre-processing techniques in this project to ensure that the textual document is more complete and efficient for data analysis.Furthermore we've also compared different Machine Learning algorithms, such as random forest and logistic regression, to see how accurate they are in terms of computation.

II. Dataset

We chose to us crime datasets for our project, hence The dataset for our project was taken from Data.world website. The website consist of 273 crime dataset. from the website we used Chris Awam's US MASS SHOOTINGS data. The data was a collection of 2 xlsx files that included A collection of indiscriminate rampages in public places resulting in four or more victims killed by the attacker. It excludes shootings stemming from more conventional crimes such as armed robbery or gang violence that occured in us from 1989 to 2019.The original data was also published in the following newspapers of USA-FiveThirtyEight: Gun Deaths in America-New York Times: Preventing Mass Shootings Like the Vegas Strip Attack Scientific American: 6 Things to Know About Mass Shootings in America Wall Street Journal (subscription): Five Things to Know About Mass Shootings in the U.S. Washington Post: America's deadliest shooting incidents are getting more deadly

III. Related Works

- Abstractive text summarization using LSTM-CNN based deep learning
  Song, S., Huang, H. amp; Ruan, T. Abstractive text summarization using LSTM-CNN based deep learning.
  Manually summarising large amounts of text is extremely difficult and time consuming for humans.So we turn to machine generated methods for it. Abstractive text summarization is such a method.Abstractive Text Summarization (ATS) is the task of creating summary sentences by combining facts from various source sentences and condensing them into a shorter representation while retaining information content and overall meaning. In this paper, they propose an LSTM-CNN architecture based on the Abstractive Text Summarization framework

(ATSDL) that can generate new sentences by exploring finer-grained fragments than mere sentences, particularly semantic phrases. The model proposed here draws on the strengths of both of the ATS and ETS models. After running multiple datasets they observed that their summarization model outperforms all the other models.

- NLP Based Latent Semantic Analysis for Legal Text Summarization
  In this paper they have proposed to make model based on Latent Semantic Approaches (LSA) to summarize the lengthy judgements of court hearing into a single document. Two summarization models were built using python. One of them used training by allowing multi-document criminal summaries to be passed to a single programme with a single execution. The other model passed civil summaries to the programme iteratively, with a new execution for each document. They were able to achieve an average ROGUE-1 score of 0.58
- Text Summarization : An overview
  The purpose of this paper is to provide an overview of Text Summarization. Text Summarization is a difficult problem to solve these days. They provided an overview of automatic text summarization. The status and state of automatic summarising have shifted dramatically over the years. Because the text summarization task has not yet been completed, and there is still much work to be done, investigated, and improved, research in this field will continue. Definitions, types, approaches, and evaluation methods have all been discussed in the paper, as have the features and techniques of existing summarization systems.
- Text Summarization with Pretrained Encoders
  In this paper they have used BERT(Bidirectional Encoder Representations from Transformers)that is considered as the latest version of pretrained language models that are available. Here they have showed how BERT can be used in text summarization and also proposed a general framework for both the abstractive and extractive models. They proposed a new fine-tuning schedule for abstractive summarization that uses different optimizers for the encoder and the decoder to alleviate the mismatch between the two. They also show how a two-stage fine-tuning approach can improve the quality of the generated summaries.
- NLP based Machine Learning Approaches for Text Summarization
  Rahul, S. Adhikari and Monika
  This paper discusses various approaches for producing summaries of large texts. Several papers have been studied for various text summarization methods that have been used in the past. The methods described in this paper primarily generate Abstractive (ABS) or Extractive (EXT) summaries of text documents. Techniques for query-based summarization are also discussed. This paper primarily discusses structured-based and semantic-based approaches to summarizing text documents. The CNN corpus, DUC2000, single and multiple text documents, and other datasets were used to test the summaries produced by these models. The authors investigated these methods, as well as their tendencies, accomplishments, past work, and future potential in text summarization and other fields. The authors said summaries produced by these methods are not always accurate. It is sometimes also irrelevant to the original document..

IV. Methodology

A. Pre-processing

Text pre-processing is the first step in the Natural Language Pre-processing (NLP) process. It is a necessary step in developing an effective machine learning algorithm that produces high-quality data and converts raw data into an useable format.
Before applying the pre-processing techniques to the data certain dispensable elements were removed from it. The features which were deemed unnecessary for summarising the text were dropped from the data frame.
The following are the various text pre-processing techniques used:

- Stopword Removal
  Stop word removal is one of the most common text pre-processing techniques used in NLP applications. The process of removing words that appear frequently in a textual document.
- Tokenization
  Tokenization is the process of splitting text document into small chunks or tokens.
  - Word Tokenizer
    Word Tokenizer it is a common prepossessing technique that divides text data into words.
  - WhitespaceTokenizer
    When it encounters a white space, the white space tokenizer separates the text into discrete tokens.
  - RegexpTokenizer A RegexpTokenizer uses a regular expression to break a string into sub-strings.
- Lemmatisation
  Natural language processing depends heavily on lemmatization. It is the process of transforming a word into its most meaningful base form. Lemmatization is context dependent. When the stemming and lemmatization techniques are compared, lemmatization outperforms stemming.

B. Term Frequency-Inverse Document Frequency

Vectorization is the process of mapping words/phrases to vectors of real valued numbers. It is used to find similarity between words, make word predictions, finding semantic similarity and so on. The vectorization technique used in this project is Term Frequency-Inverse Document Frequency (TF-IDF).

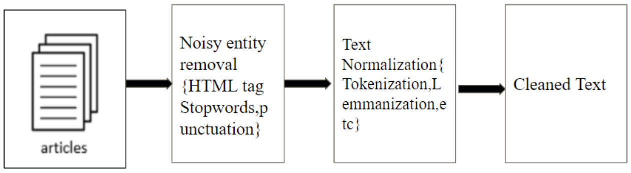

articles
Noisy entity removal {HTML tag Stopwords,punctuation}
Text Normalization{ Tokenization,Lemmanization,etc}
Cleaned Text

Fig. 1. Schematic diagram of Text Summarizing

Term Frequency and Inverse Document Frequency are shortened as TF-IDF. It is a Natural Language Processing approach that is used in a variety of NLP applications to turn words into vectors with semantic information and to give weight to unusual terms. When the TF and IDF scores are compounded, the result is TF-IDF.

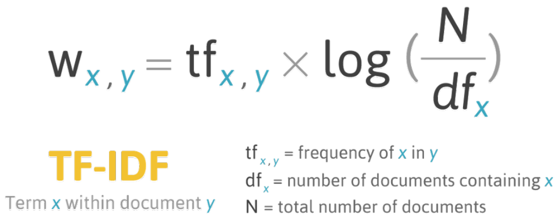

$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$
TF-IDF
Term x within document y
$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Fig. 2. TF-IFD

C. Summarization using NLTK

In the summarizer section, we first eliminated English stop words from our input text and established a dictionary named word- frequencies, which we later updated with the frequency values of the words. The variable maximum frequency is updated with the maximum values of word frequency. An iterative for loop is used to iterate through the keys in the word frequencies. After that, we tokenized the sentence and utilised the sentence score variable to update the sentence's score. To construct the final textual document summarization, we use the heapq package, which returns the most prioritized sentences as the output.

D. ML Algorithms

- K-nearest neighbors (KNN) Algorithm
  KNN is a supervised machine learning technique that can be used to solve classification and regression problems. The KNN technique is used to classify data by locating the K closest matches in training data and classifying that data into a category that is very similar to the new data. To find the closest match, a distance such as euclidean is used.
- Logistic regression Algorithm
  Logistic regression is a classification procedure that is widely used in data classification. It is used to compute or forecast the likelihood of a binary event occurring.
- Decision Tree Algorithm The purpose of this algorithm is to develop a model that outputs the value of a target variable, for which the decision tree employs the tree representation to solve the problem, with the leaf node corresponding to a class label and attributes represented on the tree's internal node.
- Random Forest Algorithm A random forest is an ensemble classifier that makes predictions based on the combination of various decision trees. It essentially fits a number of decision tree classifiers to different subsamples of the dataset. Furthermore, each tree in the forest is based on a random best subset of features.

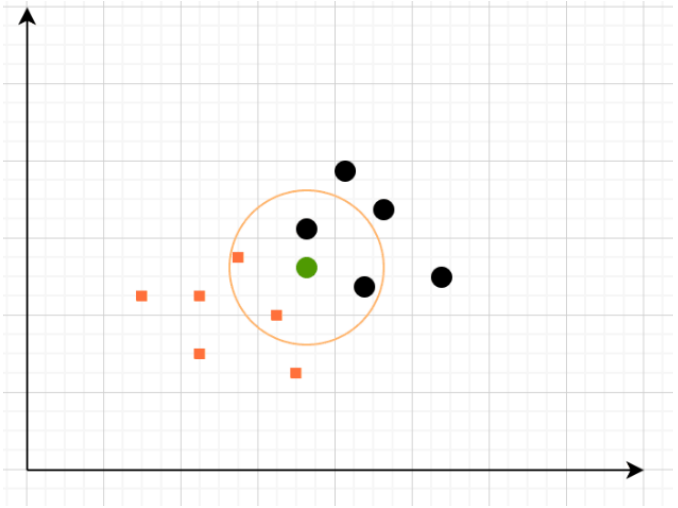


Fig. 3. KNN

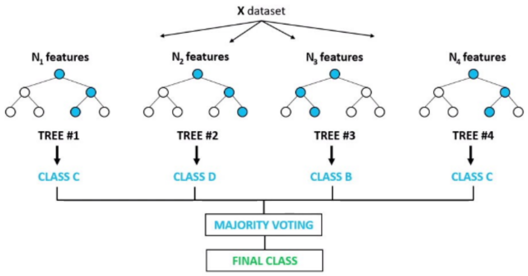


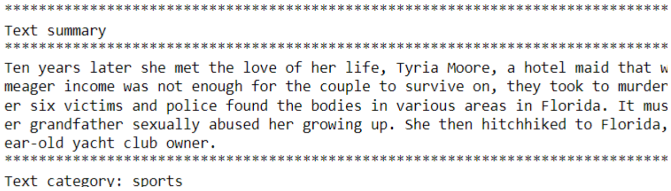
Fig. 4. Random Forest

V. Evaluation

A. Results

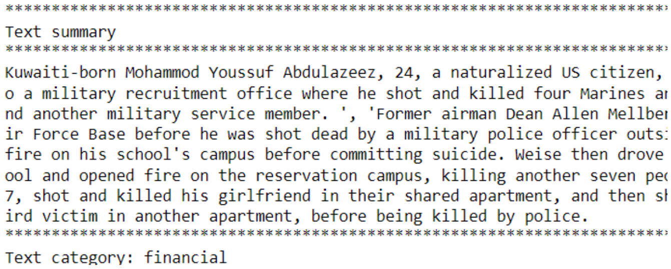
We tried various machine learning algorithms to compute computational accuracy, and a random sample of customized input was used to generate the summarised text.

Machine Learning Algorithm	Accuracy
RANDOM-FOREST CLASSIFIER	0.75
LOGISTIC REGRESSION	0.875
KNN	0.71875
Decision Tree Classifier	0.75



```
******************************************************************************
Text summary
******************************************************************************
Ten years later she met the love of her life, Tyria Moore, a hotel maid that w
meager income was not enough for the couple to survive on, they took to murder
er six victims and police found the bodies in various areas in Florida. It mus
er grandfather sexually abused her growing up. She then hitchhiked to Florida,
ear-old yacht club owner.
******************************************************************************
Text category: sports
```

Fig. 5. Custom Input 1



```
******************************************************************************
Text summary
******************************************************************************
Kuwaiti-born Mohammod Youssuf Abdulazeez, 24, a naturalized US citizen,
o a military recruitment office where he shot and killed four Marines an
nd another military service member. ', 'Former airman Dean Allen Mellber
ir Force Base before he was shot dead by a military police officer outsi
fire on his school's campus before committing suicide. Weise then drove
ool and opened fire on the reservation campus, killing another seven peo
7, shot and killed his girlfriend in their shared apartment, and then sh
ird victim in another apartment, before being killed by police.
******************************************************************************
Text category: financial
```

Fig. 6. Custom Input 1

VI. Conclusion

We were able to construct text summaries for extractive textual documents in this paper. And we were able to effectively tackle the challenges that arose as a result of the absence of a textual source. We used many machine learning algorithms to generate computational accuracy and were able to determine that logistic regression was the best accuracy prediction machine learning strategy. We were later able to successfully implement random custom input to generate text summarizing.

References

[1] Rahul, Surabhi Adhikari, and Monika, “NLP based Machine Learning Approaches for Text Summarization,” Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020).

[2] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami, Pooya Khosravyan Dehkordy, and Asghar Tajoddin, “Optimizing Text Summarization Based on Fuzzy Logic,” Seventh IEEE/ACIS International Conference on Computer and Information Science.

[3] Narendra Andhale and L.A. Bewoor, “An Overview of Text Summarization Techniques,” 2016 International Conference on Computing Communication Control and automation (ICCUBEA).

[4] Sonia Haiduc, Jairo Aponte, Laura Moreno, and Andrian Marcus, “On the Use of Automated Text Summarization Techniques for Summarizing Source Code,” 2010 17th Working Conference on Reverse Engineering.

[5] Heena A. Chopade and Dr.Meera Narvekar, ‘Hybrid Auto Text Summarization Using Deep Neural Network And Fuzzy Logic System,” Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017).

[6] N.Moratanch and S.Chitrakala, “A Survey on Extractive Text Summarization,” IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017).

[7] R. Subha Shini and V.D. Ambeth Kumar, “Recurrent Neural Network based Text Summarization Techniques by Word Sequence Generation,” Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021].

[8] Rahul, S. Adhikari and Monika, ”NLP based Machine Learning Approaches for Text Summarization,” 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.

[9] arXiv:1908.08345[cs.CL](orarXiv:1908.08345v2[cs.CL] for this version)https://doi.org/10.48550/arXiv.1908.08345

[10] Appl 78, 857–875 (2019). https://doi.org/10.1007/s11042-018-5749-3

[11] Multimed Tools Appl 78, 857–875 (2019). https://doi.org/10.1007/s11042-018-5749-3