International Journal of Recent Advances in Multidisciplinary Topics
Volume 4, Issue 3, March 2023
https://www.ijramt.com | ISSN (Online): 2582-7839

Text Categorization and Summarization

S. Sakthi Devi[1*], S. Sneha[2], S. Gururajan[3], G. Siva[4]
[1,2,3]Student, Department of Computer Technology, Bannari Amman Institute of Technology, Erode, India
[4]Assistant Professor, Department of Computer Technology, Bannari Amman Institute of Technology, Erode, India

***Abstract*: This study mainly aims on automatic text categorization and summarization. We have discussed a categorization and summarization method using machine learning techniques. Other researchers have suggested a wide variety of text categorization and summarization methods. Our paper focuses on abstractive summarization built using T5 language model. The categorization model uses logistic regression and support vector classifier algorithm. The paper emphasizes whole process of building a machine learning application from collecting datasets to simulating it on a web application. We have presented a demo sample of implementing the categorization and summarization models in a unique way separating raw and docs text from webpage text and discussed the detailed workflow. We have also discussed the process of building machine learning or deep learning models using the key technology of natural language processing. The evaluation techniques and demo working of model is also discussed.**

***Keywords*: NLP, T5, logistic regression, support vector classifier.**

1. Introduction

Today, it is commonly acknowledged that unstructured data, typically in text form such as reports, filled-out forms, emails, memos, log entries, transcripts, etc., makes up the great majority of information in any firm.

"As much as 90 percent of data is defined as unstructured data. And unstructured data is growing by 55–65 percent each year."

Companies frequently are not fully aware of its potential value because to the enormous amount of work required to manually sort through and extract information from such massive volumes. What if the document is lengthy and has only less information? What if the content we assumed was not the webpage is actually about? We end up wasting our time here. There is a need to develop machine learning algorithms that can solve these problems and help you save your time and work.

Text summarization can swiftly extract meaningful information from huge collections of documents, while text categorization may quickly predict the category of any text by combining natural language processing, statistical analysis, and machine learning approaches. Typically, this tool will go through a million words in a matter of seconds, automatically extracting subjects and revealing previously unidentified correlations and patterns. These problems comes under the domain of Artificial Intelligence. Artificial intelligence is most likely the quickest developing advancement in the world of innovation. From Siri to Alexa, self-driving vehicles, ridesharing taxis like Uber, AI makes organizations more intelligent. Artificial Intelligence may increase the efficiency of the existing economy. AI is an impersonation of human knowledge by system or machines. With AI, machines perform functions such as learning, planning, reasoning and problem solving. It is a technology that is transforming every walk of life.

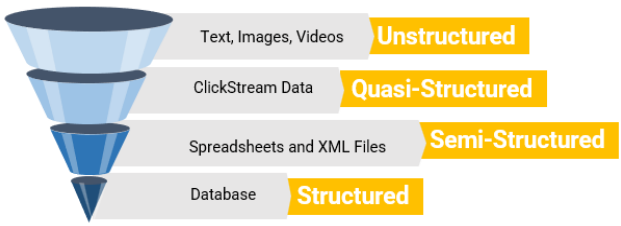

Fig. 1. Data

When we think of AI, we immediately go for robots and self driving autonomous cars. Is it just that? Now AI has already taken over the world and we see it in our everyday life . Google search engine is using machine learning algorithm to get to the right index over millions of web pages In just milliseconds So , the applications are numerous. AI is currently used in practically every industry, from marketing to healthcare. Traditionally, supervised learning techniques are used to handle the challenge of classifying text content since they assign texts to particular categories.

Text summarization and categorization can be performed by building deep learning models to give better accuracy. Machine learning, deep learning and natural language processing are parts of Artificial Intelligence. Deep learning is nothing but a subset of Machine learning and Artificial Intelligence is a superset of Machine learning. Among all these, there are 2 main applications of AI which is Natural language processing and Computer vision, the main objective of AI is to mimic human intelligence. Humans can speak, understand and communicate with each other, so 3 NLP helps computers understand, interpret and manipulate human language like SIRI and Alexa. Humans can then see and recognise objects, while computer vision makes it possible for robots to "see" and comprehend the information of digital images like pictures and movies. There are different types of text categorization. Sentiment analysis,

*Corresponding author: luwihina26@gmail.com

topic detection and language detection are few examples. Topic classification is classifying the domain or subject of the text content. For example, deciding whether the text is about “arts,” “technology,” or “sports,” or “space” is a type of topic classification.

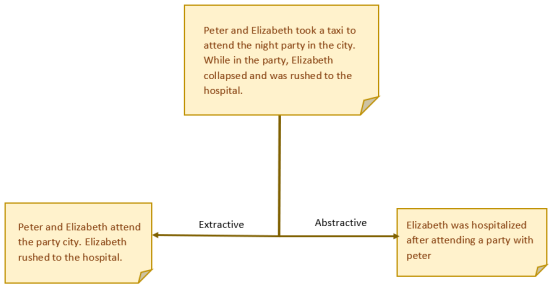

Fig. 2. Types of summarizations

A. Software Description

1) Google Colab

As the name says, colab or colaboratory is a product from google. It is free of cost. It supports many python libraries allowing you to run any python code especially used in case of building machine learning and deep learning models.

2) Visual Studio Code

Microsoft is the maker of Visual Studio Code, or VS Code. It is a web-based and desktop-compatible lightweight and potent IDE (Integrated Development Environment).

3) Pycharm

An IDE for expert Python coders is called Pycharm. JetBrains created it. It can be used with Windows, Linux, and macOS operating systems.

B. Technology Description

1) HTML, CSS and JS

HTML - Hypertext markup language as its name says it’s a markup language. It is used to give structure to web pages.

CSS - Cascading StyleSheet is to decorate or design your web pages such as adding background image, giving colors to webpage elements, aligning the content etc.

JS - Javascript is to build dynamic web pages. In this project, it is to give a popup.

2) Flask

We coded in python language to create the Flask web application framework. It is created by Armin Ronacher, the president of Pocco, a global organization of Python aficionados.

3) Beautifulsoup

It is a python package. It is used to scrape data from webpage. We used beautifulsoup to retrieve the text and image content from the web page.

4) MySQL

Based on Structured Query Language, MySQL is an open source relational database management system (RDBMS) sponsored by Oracle (SQL).

5) Ajax

A set of web development methods called Ajax uses several client-side web technologies to build asynchronous web applications.

6) Machine Learning

Artificial intelligence, which is widely defined as a machine's ability to mimic intelligent human behaviour, includes the subfield of machine learning.

7) Natural Language Processing

The field of computer science known as "natural language processing" (NLP) is more particularly the field of "artificial intelligence" (AI) that is concerned with providing computers the capacity to comprehend written and spoken words in a manner similar to that of humans.

Fig. 3. Technology stack

2. Methodology

A. Solution Description – Paragraphs & Docs

If user needs to categorize:

- He can go to categorize tab in the text analyzer and paste the text in text box or upload multiple docs.
- The text will be fed into Text categorization ML model and the output will be sent back to the website and will be displayed.
- The docs will be classified by their categories.

If user needs to Summarize:

- He can go to summarize tab and paste the text in textbox or upload document which need to be summarized.
- The text will be fed into summarization ML model and the output will be displayed similarly.

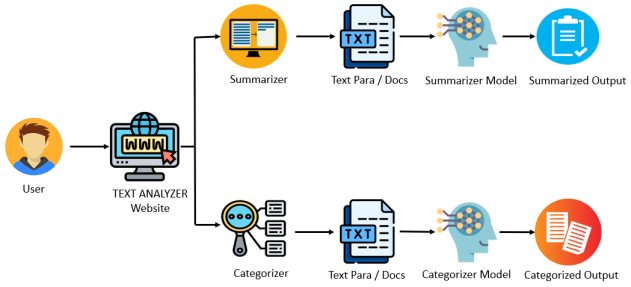

Fig. 4. Work flow for paragraphs and docs text

B. Solution Description – Webpage Texts

If user needs to categorize :

- User clicks on the URL to open to webpage.
- As soon as the webpage is clicked, the url will be fed into webscraper (beautifulsoup). The text content will be taken.

- The text content will be fed into categorization ML model and output will be displayed in popup as "This webpage has arts related text. Do you want to continue?" If yes the user can go to webpage.

If user needs to Summarize:

- Similarly the text inside URL will be fed into summarization ML model and the summarized webpage will be displayed.

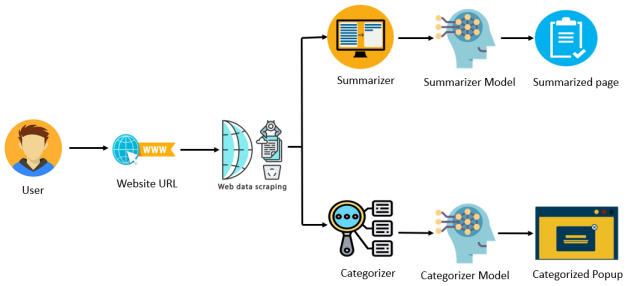

Fig. 5. Work flow for webpage texts

C. Dataset

1) Text categorization

This dataset (Document Categorization) taken from Kaggle is a collection of thousand newsgroup documents from ten different newsgroups. It is very much helpful in text classification and clustering. The ten different categories are business, entertainment, food, graphics, historical, medical, politics, space, sport and technologies. Each category contains 100 text files. For example, the category 'food' has ten text files related to food. Also, in this dataset duplicate messages have been removed.

	Title	Document	Class
0	graphics_26.txt	univers vesa driver support video includ 24 bi...	graphics
1	graphics_42.txt	articl o4fzrapdbh107h aow alan walford write h...	graphics
2	graphics_11.txt	articl 1pp991 t63 jk87377 kouhia juhana write ...	graphics
3	graphics_43.txt	interest inform stereoscop imag sun workstat p...	graphics
4	graphics_29.txt	learn adob photoshop illustr indesign less 35 ...	graphics
...	...	...	...
995	technologie_70.txt	iphon se iphon x buy refurbish phone yet appl ...	technologie
996	technologie_82.txt	game firm tough uk video game firm face test t...	technologie
997	technologie_74.txt	new consol promis big problem make game futur ...	technologie
998	technologie_87.txt	net regul blur boundari tv internet rais quest...	technologie
999	technologie_75.txt	us pirat convict first convict piraci network ...	technologie

1000 rows × 3 columns

Fig. 6. Document categorization dataset

2) Text Summarization

WikiHow is a dataset made up of more than 230,000 summary and article pairs that were taken from an online knowledge base and created by various human authors. The pieces cover a wide range of subjects and exhibit a great degree of stylistic diversity. The WikiHow dataset has three parts

Title – It has the article title

Headline – The mixing of all the headlines of all text. It helps as a reference.

Text – The mixing of all the texts/paragraphs

Here are the statistics of WikiHow dataset

The size of wikihow dataset is 230,843.

The average length of article is 579.8

The average length of summary is 62.1

The size of vocabulary is 556,461.

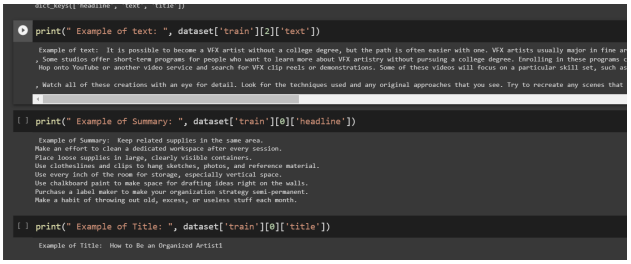

Fig. 7. Sample text in dataset

D. Model Building

Model building includes all the steps from Loading libraries/datasets to predicting the output. In this study two models need to be built. One for text Categorization (Logistic Regression) and other for text summarization (T5 Model).

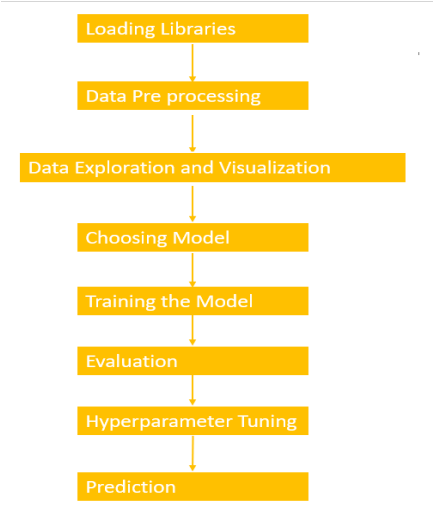

Fig. 8. Model building steps

E. Loading Libraries

- Numpy
- Pandas
- Natural Language Toolkit
- Pytorch
- Transformers
- Matplotlib

F. Data Pre processing

1) Tokenization

Tokenization is mandatory for working with text data. It is an essential step when training model using nlp. It is a process in which a piece of text (original text) is separated or splitted into smaller units called tokens (building blocks of language).

2) Lower casing

The text is made up of both uppercase and lowercase characters. As a human we know that both have same meaning. But machine don't know that they both are same.

3) Stop words removal

Stop words are the common words occurring in any text that includes articles, conjunctions, prepositions and so on. They don't give much details to the text. So, these stop words can be

filtered out before building nlp models in order to focus on the more important text.

4) Stemming

A stemming algorithm is a linguistic normalisation procedure that reduces a word's various forms to their common form, such as connection connections 17 connective.

G. Data Visualization

After loading and preprocessing the data, visualizing the data helps us to get to know the data better. That is the data is viewed by representing it by graph or charts such as pie chart, bar chart etc. The dataset is visualized.

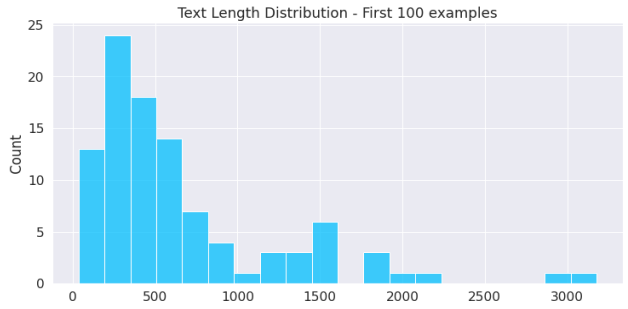

Fig. 9. Data Visualization 1

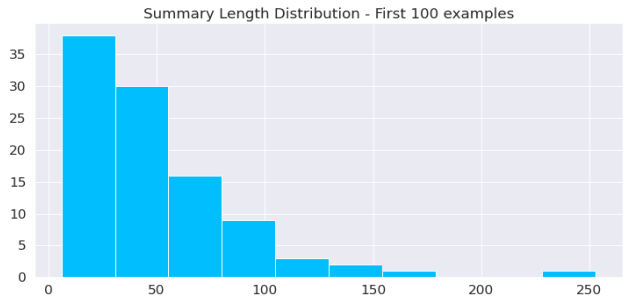

Fig. 10. Data Visualization 2

H. Choosing model

1) Text Categorization Model

- Naïve Bayes
- Decision Tree
- Random Forest
- Support Vector Classifier
- Logistic Regression

2) Text Summarization Model

- BERT
- T5

I. Training the Model

a. Dataset Splitting
b. Encoding
c. Word Vectorization or Feature Extraction

$$TF(i,j) = \frac{Term\ i\ frequency\ in\ document}{Total\ words\ in\ document\ j}$$

$$IDF(i) = log_2\left(\frac{Total\ Documents}{Documents\ with\ term\ i}\right)$$

$$TFIDF = TF * IDF$$

TF-IDF shall be used by sklearn library

J. Hyperparameter Tuning

1) Text Categorization

The three hyperparameter optimization strategy include Grid Search, Population based training and Bayesian Optimization

For the logistic regression model, hyperparameters tuning is done by finding the best parameter using Grid Search.

A grid of parameter values defined by the param grid parameter is generated exhaustively by the grid search offered by GridSearchCV. In order to "fit" the GridSearchCV instance to a dataset, all potential combinations of parameter values are considered, and the best combination is kept. This is how the typical estimator API is implemented.

In our model we got, c = 1, penalty = 'none' and solver = 'sag'

2) Text Summarization

The two types of hyperparameters include Optimizer hyperparameters and Model specific hyperparameters.

Here are some hyperparameters used for our T5 model, we used learning rate = 3e-4; adam epsilon = 1e-8; train epochs = 2 and seed = 42.

Table 1
Existing vs. Proposed system

S.No.	Existing System	Proposed System
1	In the existing system, the webpage link need to be summarized is submitted in the text box and the summarized text is displayed.	In the proposed system, summarized webpage text will be displayed in a webpage automatically using a click.
2	There is no feature to display the categorized text in popup.	The extension help us to categorize the web page text and show it in popup before the webpage opens.
3	The existing summarization website has only summarizer.	We combined the summarization and categorization in a single website.

3. Result and Discussion

A. Evaluation Metrics

- True Positive
- False Positive
- True Negative
- False Negative

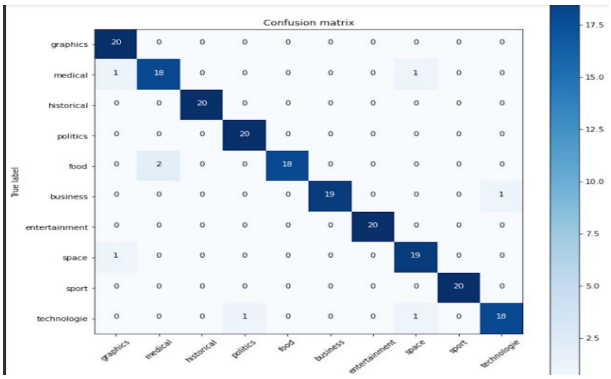

Fig. 11. Logistic regression confusion matrix

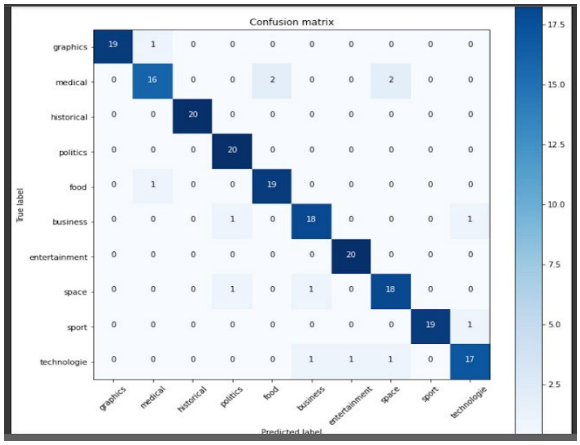

Fig. 12. SVC confusion matrix

Precision:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Precision = \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Recall:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive}$$

F1 Score:

$$F1 = 2\ x\ \frac{Precision * Recall}{Precision + Recall}$$

Table 2
Result

Problem	Chosen Model	Reason
Text Categorization	Logistic Regression	Gives better accuracy compared to other algorithms.
Text Summarization	T5	Able to process large amount of texts

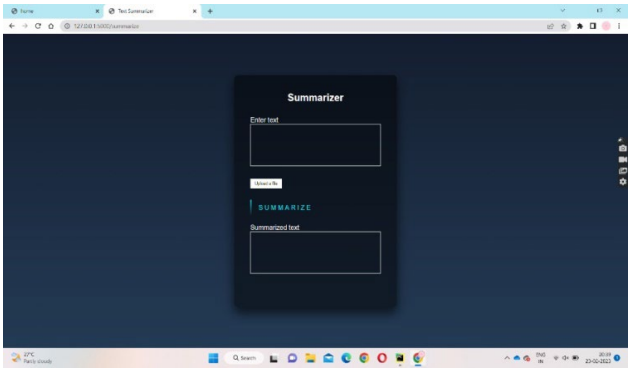

Fig. 13. Text summarizer page

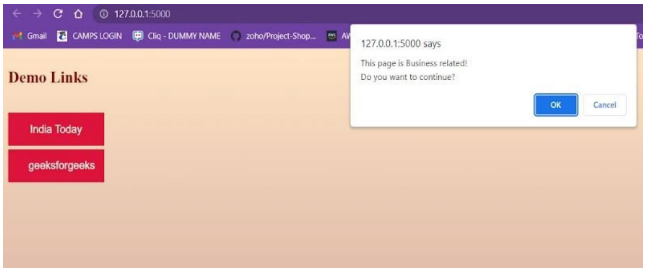

Fig. 14. Text categorizer output

4. Conclusion

Text Categorization and Summarization helps in analysing textual data. In this study we focused on various machine learning and deep learning models to categorize and summarize the text. The accuracy of each model is analyzed. And through this study we made conclusions to use the Logistic Regression, T5 model for categorization and summarization for its best accuracy. This model was integrated with backend using flask and a pop-up message is shown as soon as the link is clicked and before the user navigates into the page. This shall be developed as extension. Also, the model is integrated with Text Analyzer webpage to categorize and summarize the raw texts and documents. Future plans of this study include summarizing the video and images and also displaying the webpage in summarized version including image in it.

References

[1] Divya Khyani (2020), An Interpretation of Lemmatization and Stemming in Natural Language Processing.
[2] Zhao Jianqiang and Gui Xiaolin, (2017), Comparison Study on Text Pre-processing Techniques.
[3] Saggion and Lapalme (2000), Concept Identification and Presentation in the Context of Technical Text Summarization.
[4] Altınel B, Ganiz MC & Diri B., (2015), A corpus-based semantic kernel for text categorization by employing mean values of terms.
[5] Demidova L. (2017), Intellectual methods to improvement of the classification decisions quality on the base of the SVM classifier.
[6] Yang L., (2014), Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization.