Project Proposal

Project : Article Summarization and Categorization

Team Members :

Divya Geethanjali Birudharaju
Jaya Prakash Reddy Gade
Ayyappa Ankishetty

Goals and Objectives :

- **Motivation :**
  [2] The motivation behind this project is, we find few articles on Instagram (captions of post), Blog or any social platform but it would be difficult for us to find what is in the article before reading it. Sometimes, after seeing the length of the article, we might loose interest and stop reading it. But, if the article is about one of our interests, we might miss reading it. So, if there is a tool which categorizes and summarizes articles, it would be a time saving for most of the people. We can read the article if it is about our interested topic. We can skip reading the article if it is not in our interested category. [2]

- **Significance :** [1]
  - Summarization helps individuals to ignore unimportant information in the text.
  - Categorization helps people to identify what is the article about.
  - Summarization and Categorization both can help lot of people, especially students to read the articles which are related to the topic they want to learn about.

- **Objectives :**
  1. To create a model which summarizes and categorize articles.
  2. Train the model with the existing data of article summarization and categorization
  3. To achieve a model with a good accuracy.
  4. To create a user-friendly UI which is connected to the backend.

- **Features :**
  - Text Classification
  - Text Summarization
  - Tokenization
  - Special Character removal
  - Stop word removal

Increment 1 :

Related Work :

- We have done more research to decide which is best technique for summarization. [3]
- As we are dealing with articles, we had difficulty in finding a dataset which would classify our articles. Finally, we have decided BBC news classification would be best for our project as it has more categories.

Dataset : (https://www.kaggle.com/c/learn-ai-bbc)

As we mentioned earlier, the dataset we are using is BBC news classification. It contains 2225 articles, each labeled in one of five categories: business, entertainment, politics, sport, or technology.

```
[17] from tensorflow import keras
     layers = keras.layers
     models = keras.models
```

```
data = pd.read_csv("bbc-text.csv")
data.head()
```

	category	text
0	tech	tv future in the hands of viewers with home th...
1	business	worldcom boss left books alone former worldc...
2	sport	tigers wary of farrell gamble leicester say ...
3	sport	yeading face newcastle in fa cup premiership s...
4	entertainment	ocean s twelve raids box office ocean s twelve...

```
data['category'].value_counts()
```

```
sport            511
business         510
politics         417
tech             401
entertainment    386
Name: category, dtype: int64
```

```
[20] train_size = int(len(data) * .8)
     print ("Train size: %d" % train_size)
     print ("Test size: %d" % (len(data) - train_size))

     def train_test_split(data, train_size):
         train = data[:train_size]
         test = data[train_size:]
         return train, test
```

```
Train size: 1780
Test size: 445
```

Detail Design of Features :

- **Text Summarization :** For text summarization, we are using extractive summarization using cosine similarity technique.

  Extractive summarization extracts information from the original text that is identical to the original content. In other words, rather than producing a unique summary based on the entire document, it will rank each sentence in the document against all of the others based on the sentence's explanatory power.

- **Text Classification :** For text classification, we are training model with BBC news classification dataset.

- **Tokenization :** For tokenization, we are using in-build nltk tokenizer.

  Tokenization is the process of separating the initial source content into element tokens. Tokenization is always the first stage in successive NLP processes.

Analysis : When we are analyzing the model to be used for text summarization, we found that extractive gives more accurate results than abstractive and faster than abstractive.

Implementation :

- When the article is processed to the model, it first tokenizes the sentence and removes stop words and gives summary of the article using extractive summarization. We will train the model with BBC news classification.
- Our summarized text from step 1 is used by the model to predict the article falls under which category.

Preliminary Results :

```
Epoch 1/2
51/51 [==============================] - 1s 6ms/step - loss: 0.4312 - accuracy: 0.8864 - val_loss: 0.1490 - val_accuracy: 0.9382
Epoch 2/2
51/51 [==============================] - 0s 4ms/step - loss: 0.0548 - accuracy: 0.9888 - val_loss: 0.1180 - val_accuracy: 0.9663
14/14 [==============================] - 0s 2ms/step - loss: 0.1443 - accuracy: 0.9551
Test loss: 0.14432784914970398
Test accuracy: 0.9550561904907227
```

```
Predicted label: tech
Summary:
ultimate game award for doom 3 sci-fi shooter doom 3 has blasted away the competition at a major games ceremony the golden joystick awards . it was the only title to win twice winning ultimate game of the year and best pc game at t

Predicted label: business
Summary:
algeria hit by further gas riots algeria suffered a weekend of violent protests against government plans to raise gas prices local press reports . butane gas and fuel oil are used as the main source of fuel to heat homes and cook f

Predicted label: tech
Summary:
fast lifts rise into record books two high-speed lifts at the world s tallest building have been officially recognised as the planet s fastest . the lifts take only 30 seconds to whisk passengers to the top of the 508m tall tfc 101

Predicted label: entertainment
Summary:
muslim group attacks tv drama 24 a british muslim group has criticised the new series of us drama 24 which is about to be aired on sky one claiming it portrays islam unfairly . the muslim council of britain has complained to broadc
```

Project Management :

- **Implementation Status Report :**

  - **Work Completed :**

    - **Description :** We have completed a task to generate summarization of article and we have trained our model to classify the text category.

    - **Responsibility :**
      - Text summarization : Jaya Prakash Reddy
      - Text Classification and training the model : Divya Geethanjali and Ayyappa
      - Documentation : Divya Geethanjali, Jaya Prakash and Ayyappa.

    - **Contribution :**
      - Divya Geethanjali Birudharaju : 33%
      - Jaya Prakash Reddy Gade : 33.5%
      - Ayyappa Ankishetty : 33.5%

  - **Works to be Completed :**

    - **Description :**
      a. We have trained the model with less dataset. We want to train the model with more dataset.
      b. We also want to implement abstractive summarization and we want to use different models to get the text categorization and summarization.
      c. We want to build a UI which is user friendly and can be easily used by everyone.

    - **Responsibility :**
      - Dataset Creation : Ayyappa Ankishetty
      - Training the model : Divya Geethanjali Birudharaju
      - User Interface : Jaya Prakash Reddy Gade
      - Documentation : Divya Geethanjali, Jaya Prakash and Ayyappa.

    - **Contribution :**
      - Divya Geethanjali : 33%
      - Jaya Prakash Reddy Gade : 34%
      - Ayyappa Ankishetty : 33%

References :

[1] Extractive Text Summarization Using Sentence Ranking by J.N. Madhuri; R. Ganesh Kumar.

(https://ieeexplore.ieee.org/abstract/document/8817040)

[2] Extractive Text Summarization Using Recent Approaches by Avaneesh Kumar Yadav* , Ashish Kumar Maurya, Ranvijay, Rama Shankar Yadav.

(https://www.iieta.org/journals/isi/paper/10.18280/isi.260112)

[3] Review of automatic text summarization techniques & methods by Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, De Rosal Ignatius Moses Setiadi.
(https://www.sciencedirect.com/science/article/pii/S1319157820303712)

GitHub Link : https://github.com/jayaprakashreddy007/NLP/tree/main/Project/Documentaion

Increment 2:

- **Introduction :**

Summarization is the process of compressing a block of content into a simpler form, minimizing the characteristics of the input text whereas conserving crucial information - based aspects and information definition. Because mechanical summary is a time-consuming and usually tedious task, automating the task is getting more popular and thus serves as a great incentive for scientific work. In deep - learning and generative grammar processing, text summarization is a frequent issue (NLP).

- **Background :**

Many such papers on text summarization have been reviewed in this article, as shown below.

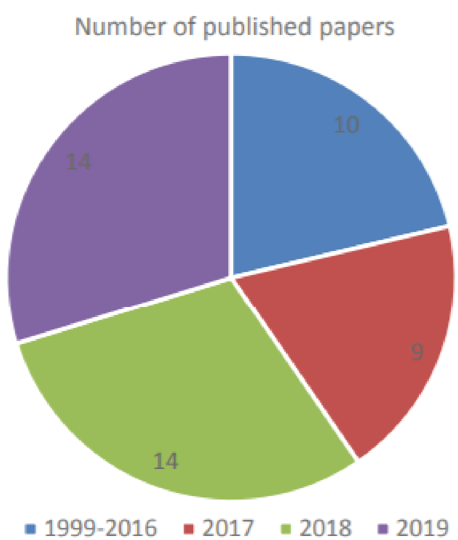


Figure 1.1: Number of Published Papers

Because there is a massive amount of data surfacing digitally, developing a punctuate procedure to immediately shorten long texts while retaining the main idea is critical [1]. Summarization also aids in reducing the amount of time spent reading, speeding up the search for information, and obtaining the most information on a single topic [2]. Text summarization is a fascinating learning topic that has gained a lot of attention recently. As research progresses, they hope to see a breakthrough that will affect this by providing a timely method of summarizing long texts [1].

This paper demonstrates a coded process for data analysis that relies on vague fundamentals on a variety of obtained data to discover one of most important data in the responded to the survey texts; the resulting deliverable synopses of these writings are compared to benchmark synopses created by website specialists. Unlike other composing approaches in the literature, this technique encapsulates writings by looking for cross - correlation showcases to reduce level of complexity and hence amount of blurry benchmarks used for summarization. As a result, the suggested technique for information summaries with a small handful of rule base can benefit the development and usage of subsequent intelligent machines capable of evaluating writing [3].

The participant's goal in summing up any content is to obtain goal and brief ideas about it. It differs from person to person for whom the text was written, as well as the length of the original document, because there are lots of reasons why a person should encapsulate a text: Overview for research, in which the worker operates to try writing critical things in the message in order to successfully gain knowledge the prerequisites. Summarize the text aids viewers in reading and comprehending it. The executive summary can also be used to review and synthesize the data written the about topic in documents and scientific papers. When summing up any document, there's several factors in play: Indicate all essential facts and concepts in the writing and leave out any irrelevant details. Need not replicate any information; instead, remove all duplicates. Simplified terms should be used in place of more difficult terms.[4]

Based on our background work, we decided to perform our summarization on two methods and categorization on seven methods. User can choose the summarization and categorization technique.

- **Our Model :**

  We have a total of 10 models in our project. 7 for text categorization and 3 for text summarization.

  - **Text Summarization Models :**

    We have used 3 models for text summarization.

    - **TextRank Algorithm Model :**

      TextRank's origins can be traced back to Google's PageRank, which is used to rank webpages for online search results.

      TextRank employs the same logic as PageRank, and with a few slight variations:

      1. Instead of Internet pages, content sentences are used.

2. The cosine similarity for index is packed with similarity attributes among phrases A & B instead of 1/total links from Site B to A, that can be estimated utilizing cosine distance.

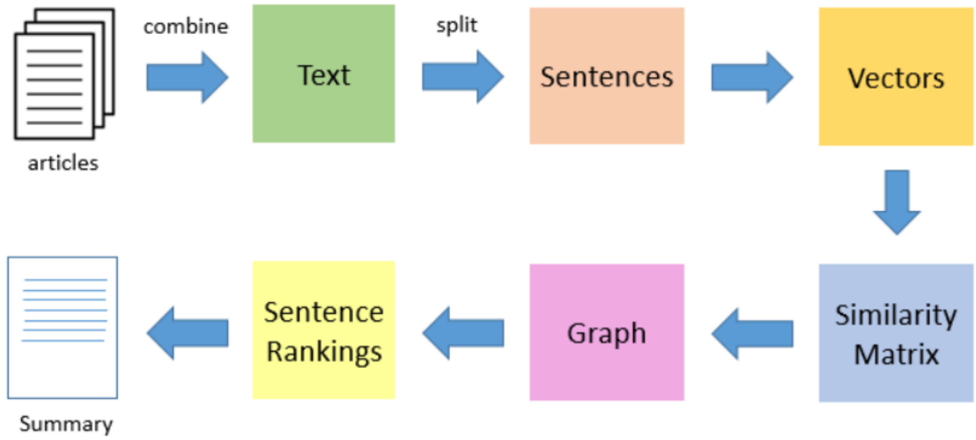


Figure 2.1: TextRank Algorithm

➢ **Gensim Model :**

Word2Vec, as outlined by TensorFlow, is indeed a prototype for gaining knowledge vector representations known as "word embeddings" developed by Mikolov et al .Word2Vec is composed of three major construction blocks. Both of them would just be thoroughly examined:

1. Frame of reference Builder Vocabulary Developer
2. A two-layer neural network
3. Gensim is a Python library for creating word2vec.

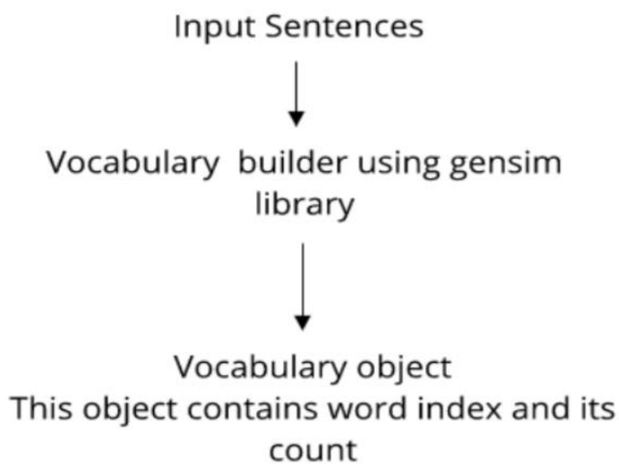


Figure 1.1: Basic building block of Word2Vec model.

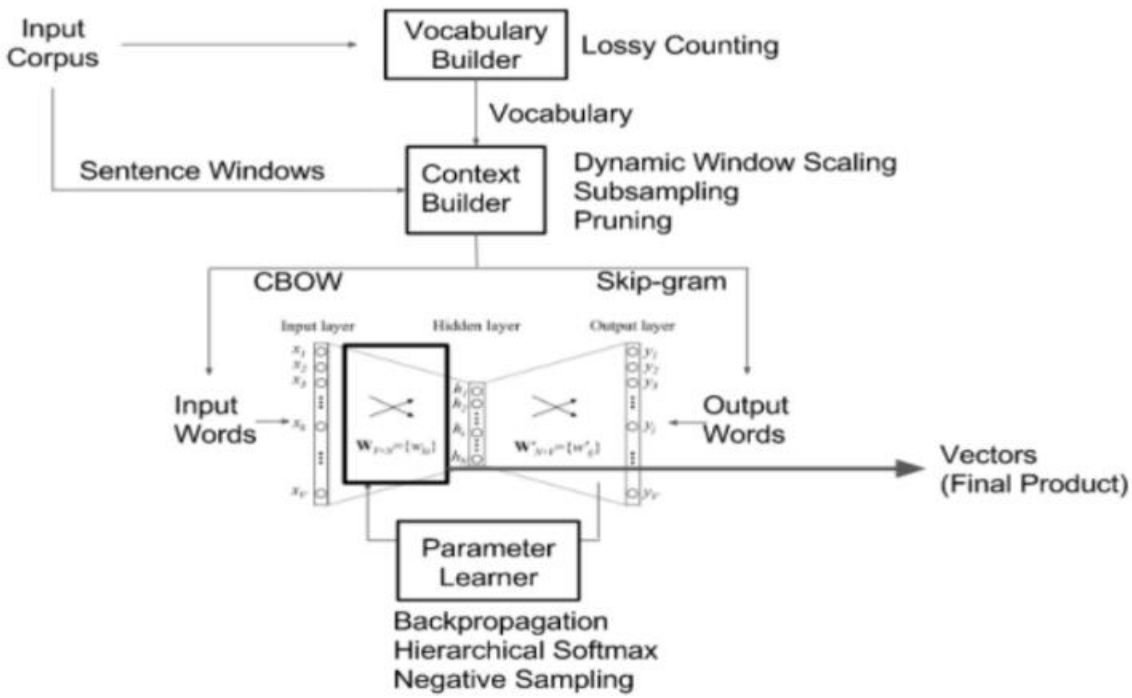
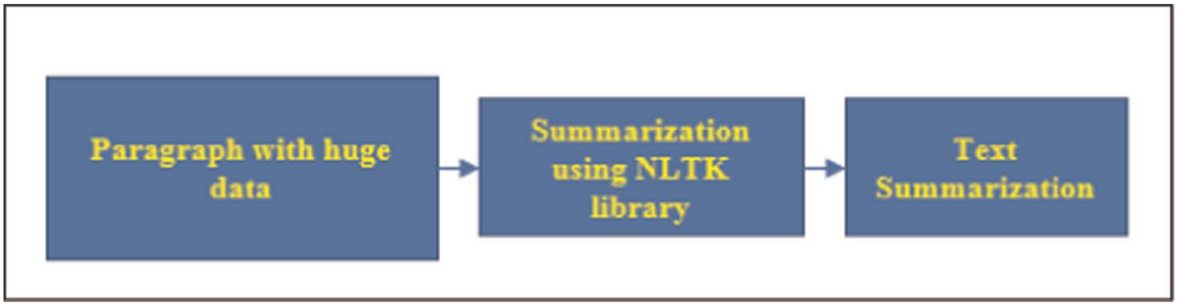


Figure 1.2: Genism Architecture

- **NLTK Summarizer :**

  NLTK summarizer is a in-built library.



- **Text Categorization Models :**

  Cleaning of text since most hyperlinks or documents contain lots of noise, and then we'll throw the cleaned text to the TF-IDF vectorizer to start preparing it for the classification algorithm. The following illustration shows the process of categorizing:

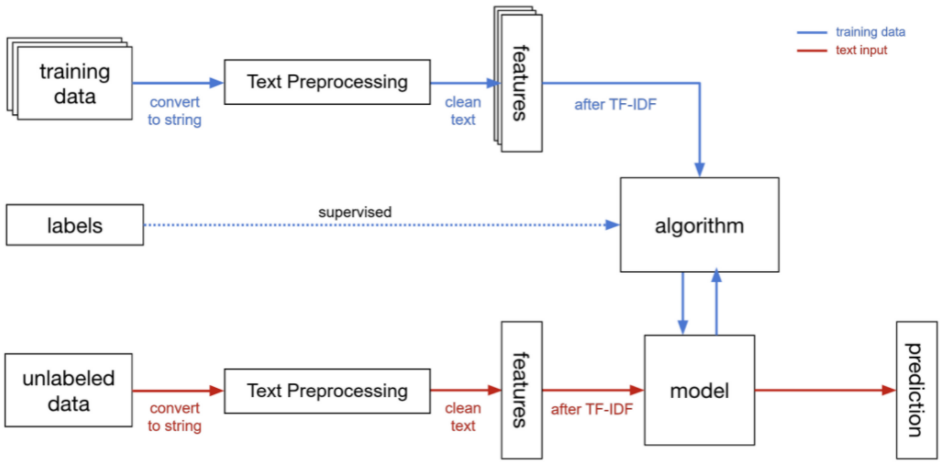


Figure 1.3: TF-IDF Architecture

We trained six various versions, because all of those models are built-in to the sklearn library, and all these models provide me with a highly accurate score for classification process.

1. Random Forest :

Random forests are an iterative procedure for text classification. Random forest is a versatile, user-friendly machine learning algorithm which generates excellent results almost all of the time with or without hyper-parameter intonation. For its simplicity and variety, it has been one of the most popular algorithms (it can be used for both classification and regression tasks). T. Kam Ho was also the first to introduce this technique, which has used tree branches in concurrently, in 1995. L. Breiman eventually developed the above method in 1999. Random Forest can also be termed as Random Decision Forests.

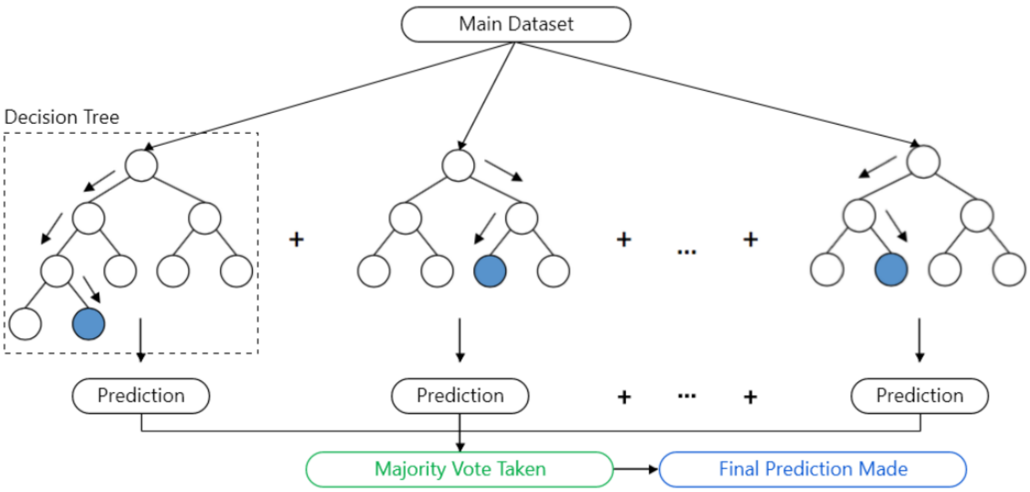


Figure 1.4: Random Forest Architecture

2. Ridge Classifier :

When the amount of predictors in a set higher than the number of findings, or when a data set has collinearity, ridge regression is used to generate a precise model. Tikhonov's method is similar to ridge regression, but Tikhonov's has a bigger sample. This can ability to design even if your data set contains a high degree of statistical noise.

3. K-Nearest Neighbours :

The k-nearest neighbor's algorithm is a semi classification algorithm. In recent years, this technique is being used in NLP as a text classifier in a wide range of research. The K-NN method implies resemblance in between new specific instance and existing instances and places the new case in the category which is most compatible with the existing categories.

The K-NN method stores all data available as well as categorizes new information points groups of similar. This implies because when new information is produced, it can even be quickly categorized into a well-suited classification using the K- NN algorithm.

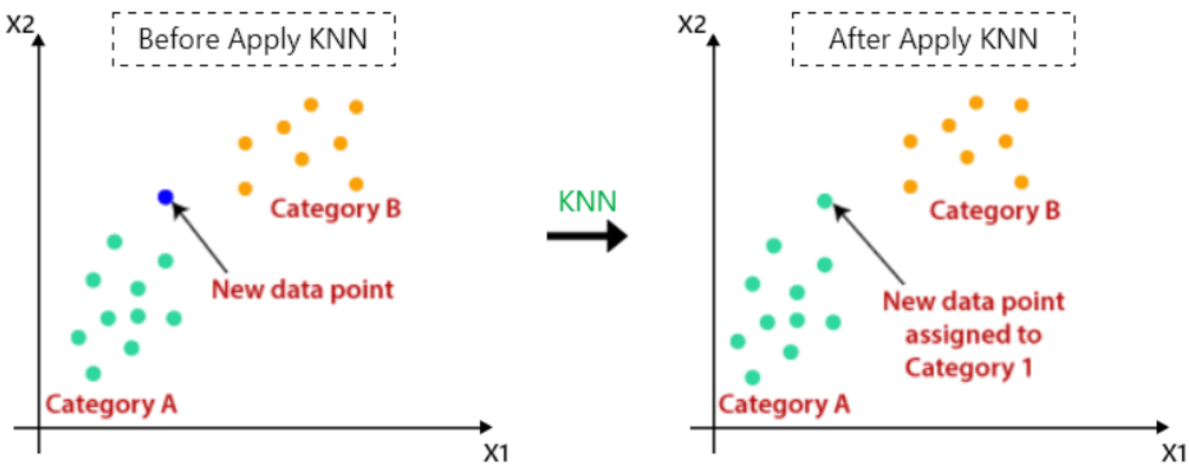


Figure 1.5: KNN Algorithm

4. Gaussian Naïve Bayes :

The assumption of independence hypotheses in between characteristics are one of the assertions used throughout the (NLP). Such classifiers make the assumption that the value of one characteristic is self sufficient of both the value of any other feature. Naive Bayes Classifier is trained extremely effective in a supervised learning environment.

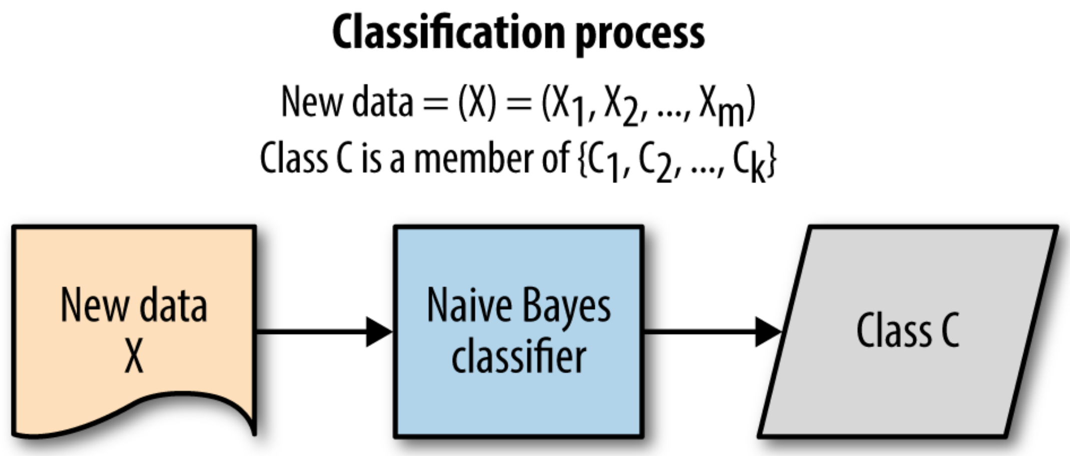


Figure 1.6: Gaussian Naïve Bayes Algorithm

5. Decision Tree :

A decision tree was one of the first text and different data mining techniques. Decision tree classifiers are used effectively in a wide range of classification tasks. The above technology's framework involves a data structure space breakdown. D. Morgan suggested and JR. Quinlan established the decision tree like a classification task. The primary idea of building trees different attributes of such pieces of data, however the task is deciding that what qualities belong just at parent stage or which belong just at child level. De Mantras proposed analytical models for selecting features in trees to fix this problem.

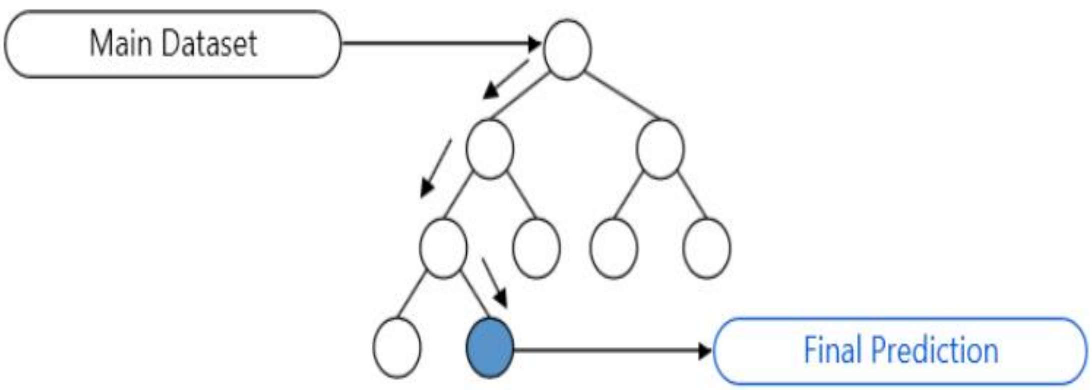


Figure 1.7: Decision Tree Algorithm

6. Logistic Regression :

When the variable is categorical, logistic regression is indeed the optimal model analysis to use. Logistic regression, like every regression evaluations, is a predictive analysis. Logistic regression is a statistical analysis technique which is used to describe the correlation between the dependent parameter and one or more independent variables.

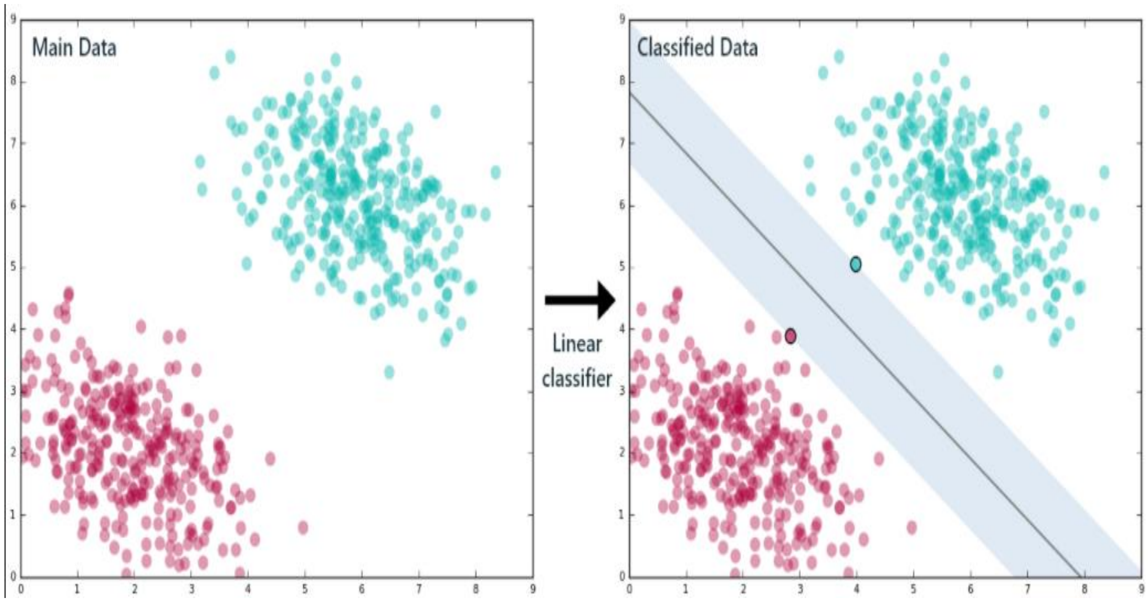


Figure 1.8: Logistic Regression Algorithm

7. **Keras Tokenizer :**

Keras' Tokenizer class is often used to vectorize a text corpus. Thus every content input is transformed into a numeric series or a variable with a correlation for every token inside the form of binary value systems for this purpose.

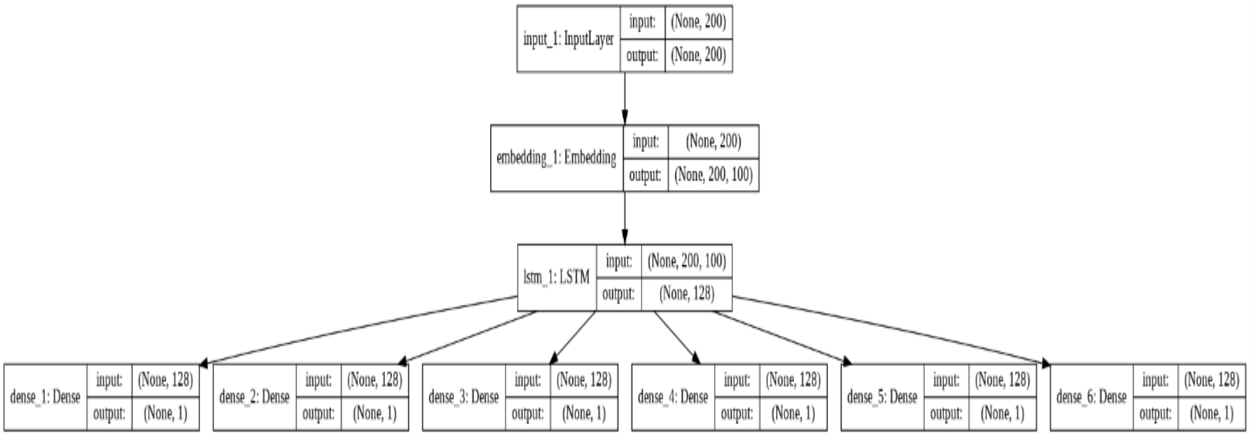
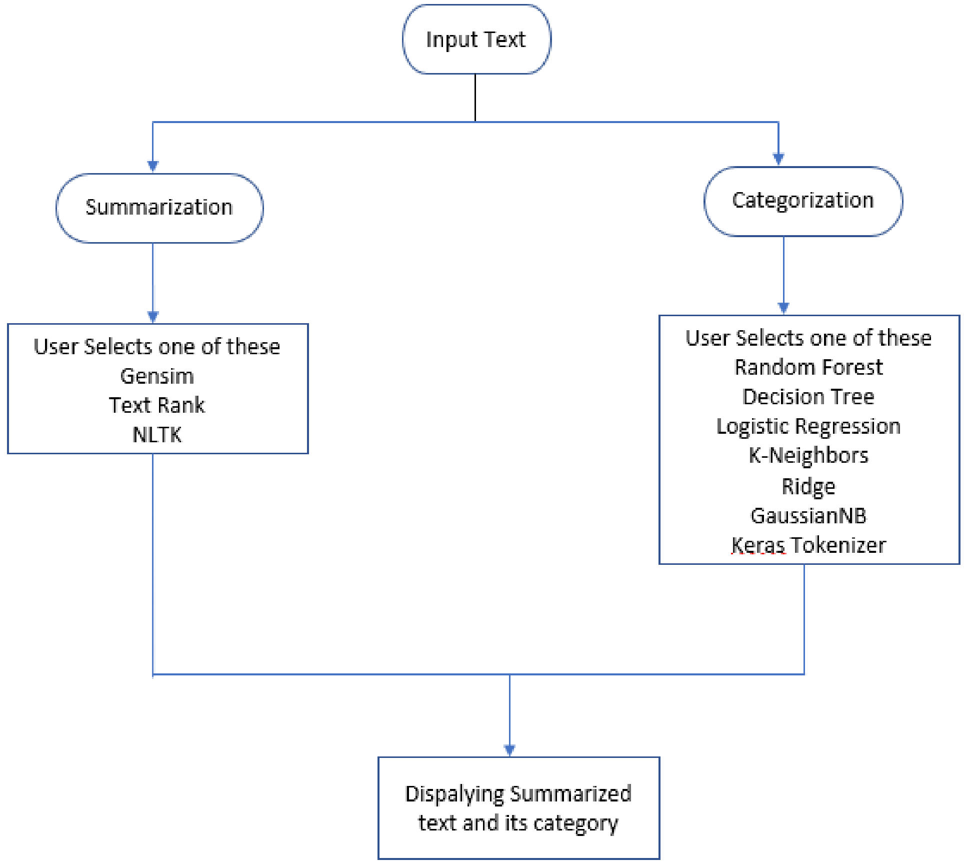


Figure 1.9: Keras Tokenizer Algorithm

- **Workflow Diagram with explanation :**



- **Dataset :**

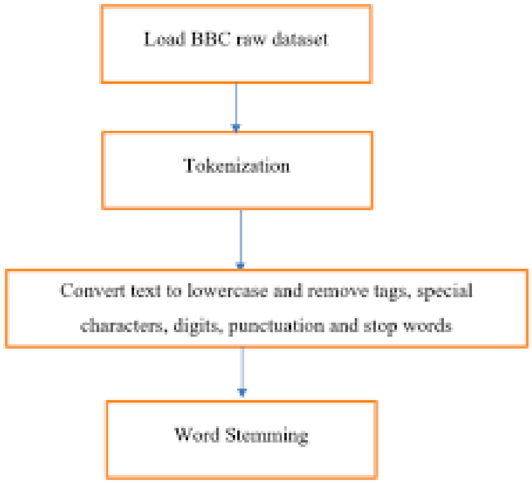
The dataset we are using is BBC news classification. It contains 2225 articles, each labeled in one of five categories: business, entertainment, politics, sport, or technology.

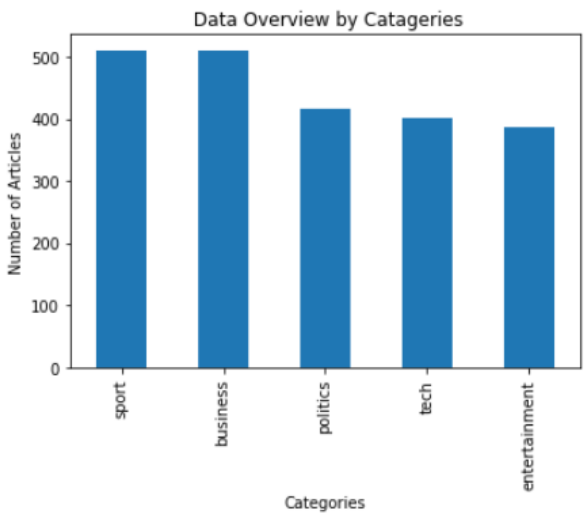
	Text	Category
0	Monsanto fined $1.5m for bribery\n\nThe US agr...	business
1	GM, Ford cut output as sales fall\n\nUS car fi...	business
2	Wal-Mart fights back at accusers\n\nTwo big US...	business
3	Rank 'set to sell off film unit'\n\nLeisure gr...	business
4	Georgia plans hidden asset pardon\n\nGeorgia i...	business
...	...	...
2219	Bond game fails to shake or stir\n\nFor gaming...	tech
2220	Europe backs digital TV lifestyle\n\nHow peopl...	tech
2221	The future in your pocket\n\nIf you are a geek...	tech
2222	California sets fines for spyware\n\nThe maker...	tech
2223	US duo in first spam conviction\n\nA brother a...	tech

2224 rows × 2 columns

- **Detail Design of Features with Diagram :**

We don't have many features for our dataset. Our dataset is folder which has subfolders. We have assessed the file and stored the information in dataframe with 3 columns Article ID, text and Category.



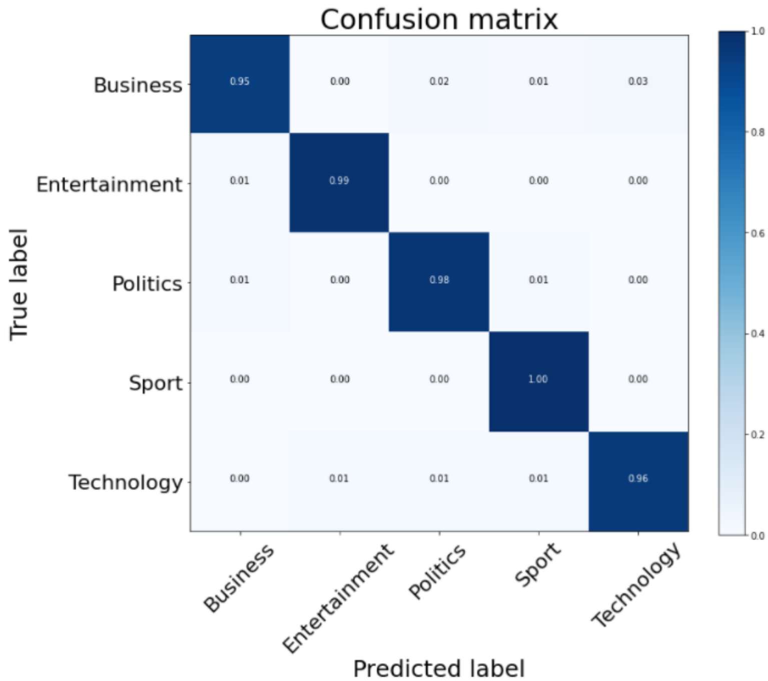


- **Model Accuracies :**

The condition to be exact or accurate is referred to as accuracy. So, accuracy is a measure to which you are nearer to a given figure. This is among the most essential and extensively utilized performance evaluation measures for categorization in machine learning.

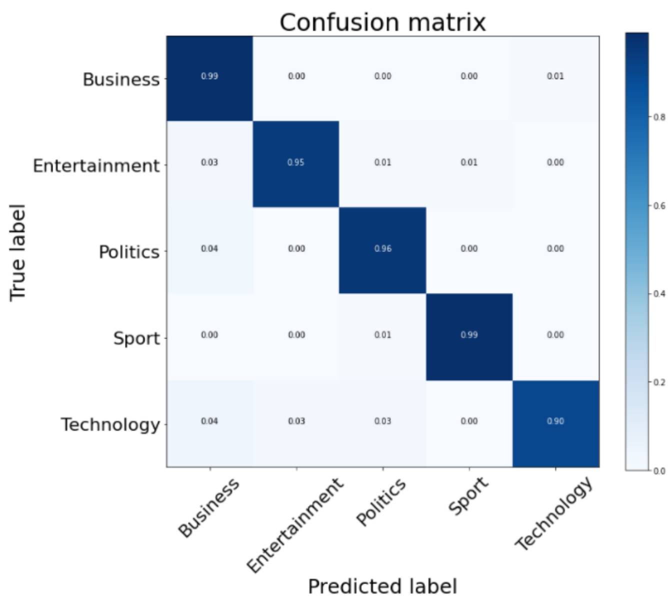
A confusion matrix, often called as an error matrix, is a summarized tables that used evaluate a categorization model's performance. The quantity of right and unsuccessful predictions is totaled and decomposed by class using score values.

- Ridge Classifier :



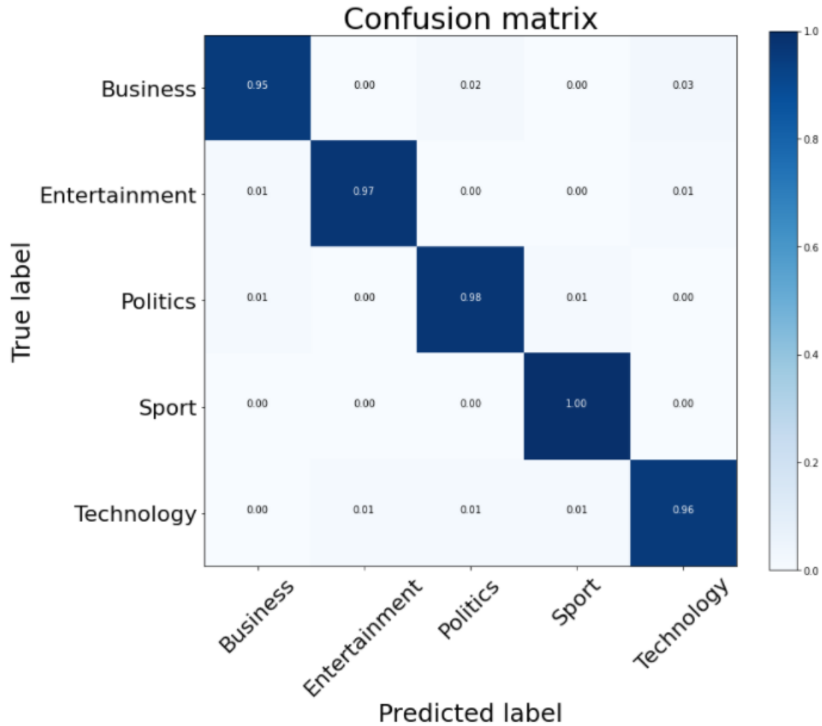

```
Accuracy on test:  0.9730337078651685
```

- Random Forest Classifier :



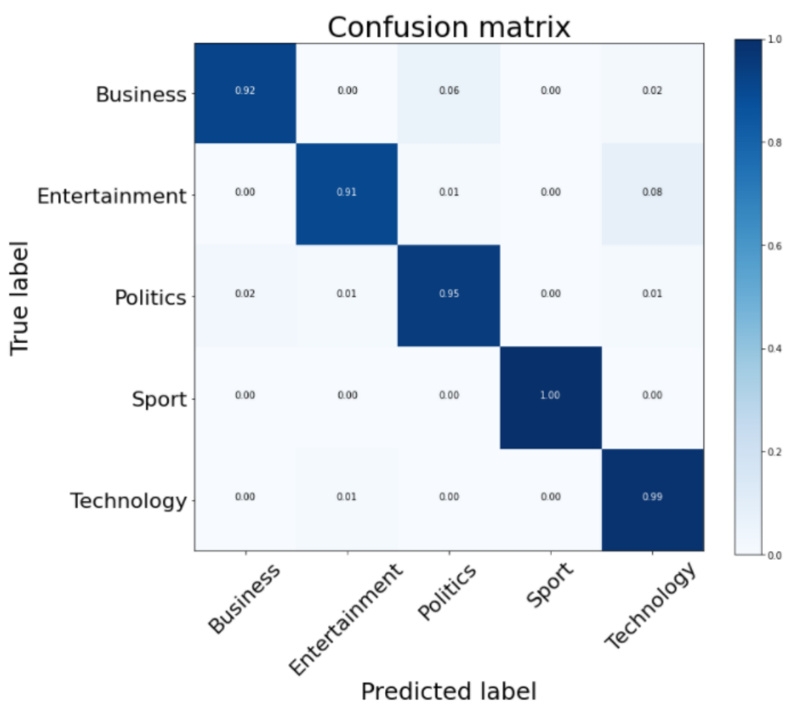

```
Accuracy on test:  0.9640449438202248
```

- Logistic Regression Classifier :



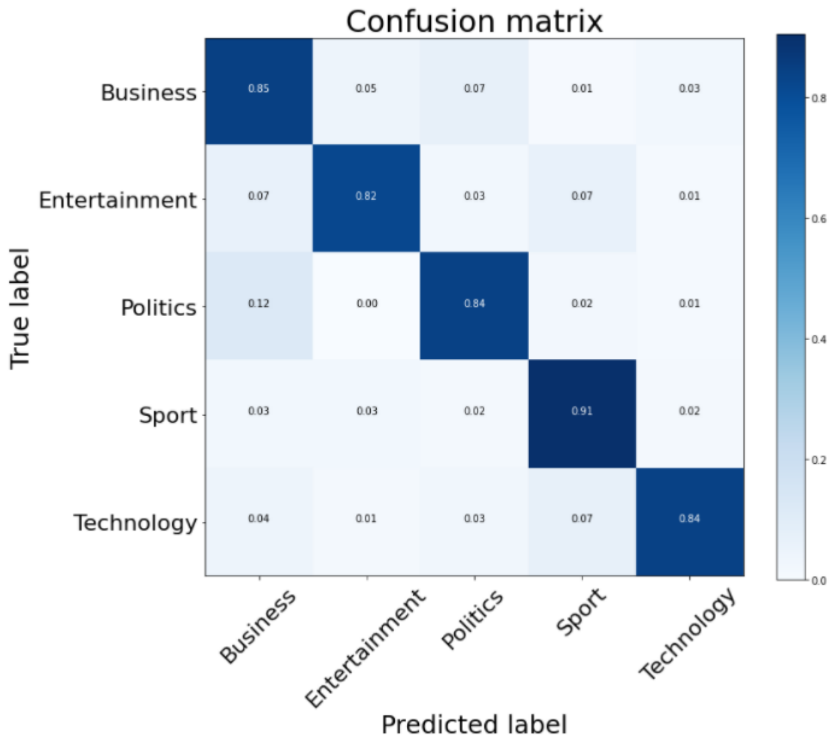

```
Accuracy on test:  0.9730337078651685
```

- KNN classifier :



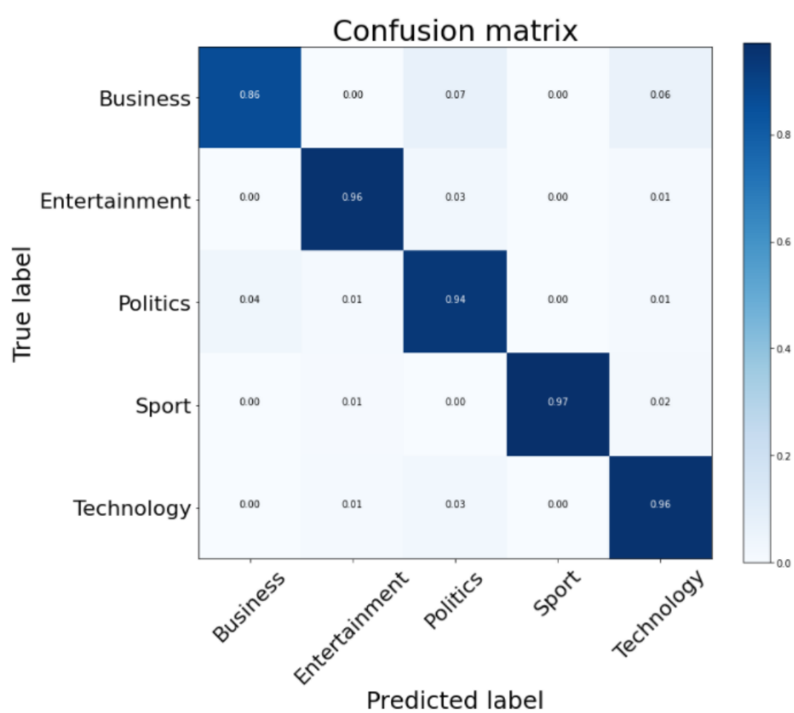

```
Accuracy on test:  0.952808988764045
```

- Decision Tree classifier :



Accuracy on test: 0.8561797752808988

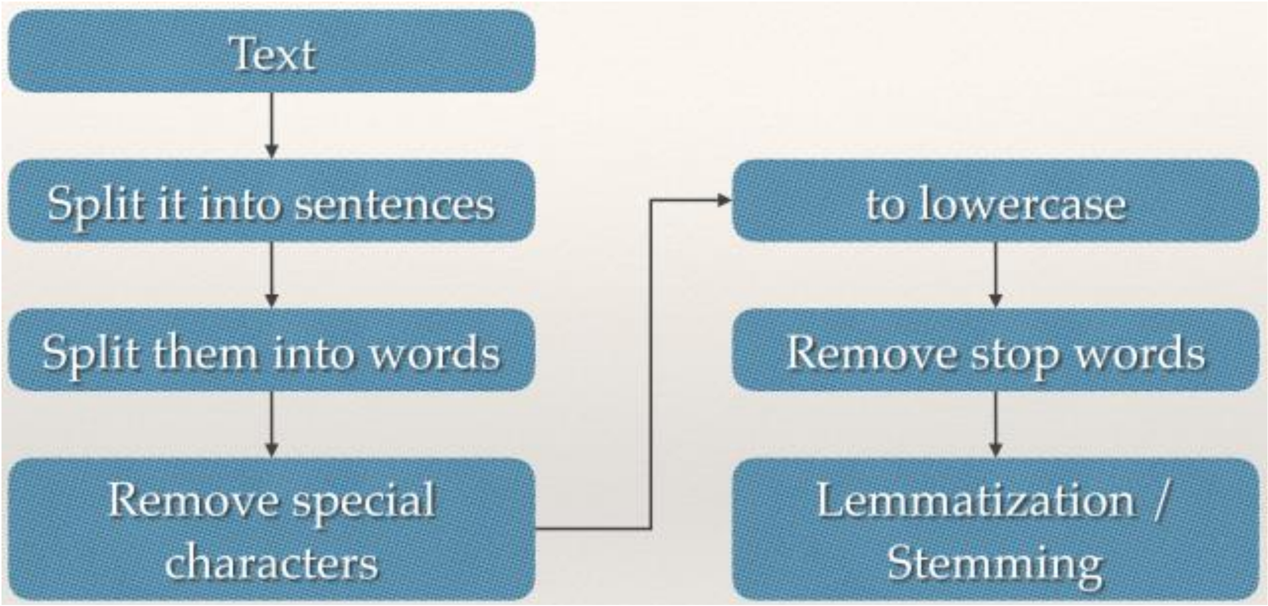
- Gaussian Naïve Bayes classifier :



Accuracy on test: 0.9348314606741573

- **Analysis Of Data :**

  - **Data Preprocessing :**



Most text and documents in Natural Language Processing (NLP) contain numerous phrases that really are useless for text categorization, including stop - words, stemming, lemmatization and so on.

Noise and unneeded characteristics can have a negative effect on the overall performance in so many methodologies, such as conventional statistical learning techniques. As a result, eliminating these features is critical. To the input text, we will take several steps, to clean the text.

➢ **Links Deletion :**

It will eliminate all links as from input text and replace them with the following:

- Identifies http protocols such as http:// or https://
- After http protocols, match optional whitespaces.
- Matches may or may not include the www.
- Whitespaces in links can be matched if desired.
- 0 or more of one or more-word characters followed by a period are matched.
- 0 or more of one or more words (or a dash or space) followed by
- Any remaining path at the end of the url, followed by an optional conclusion.
- Matches query parameters at the end (even with white spaces, etc).

- **Fixing Word Length :**

  When characters are incorrectly repeated, they cause word lengthening. English words can only have two repeated characters, such as the words wood and school in English. Additional characters must be ripped off; otherwise, we risk including misleading information. Substitute the special letters with another letter.

- **Remove Unwanted Symbols and Stopwords :**

  Another issue of text cleaning as a pre-processing step is noise removal. Text documents generally contains characters like punctuations or special characters and they are not necessary for text mining or classification purposes. Although punctuation is critical to understand the meaning of the sentence, but it can affect the classification algorithms negatively.

  - Example of bad symbols /(){}\[\]|@âÂ,;\?\'\"*…؛:-+\!&‘’–؟-
  - Stopwords like prepositions and hyphens words. for example and, in, or, etc

- **Stemming :**

  Text stemming is the process of modifying a word to acquire variations through the use of different language procedures such as addition of affixes.

- **Lower Case :**

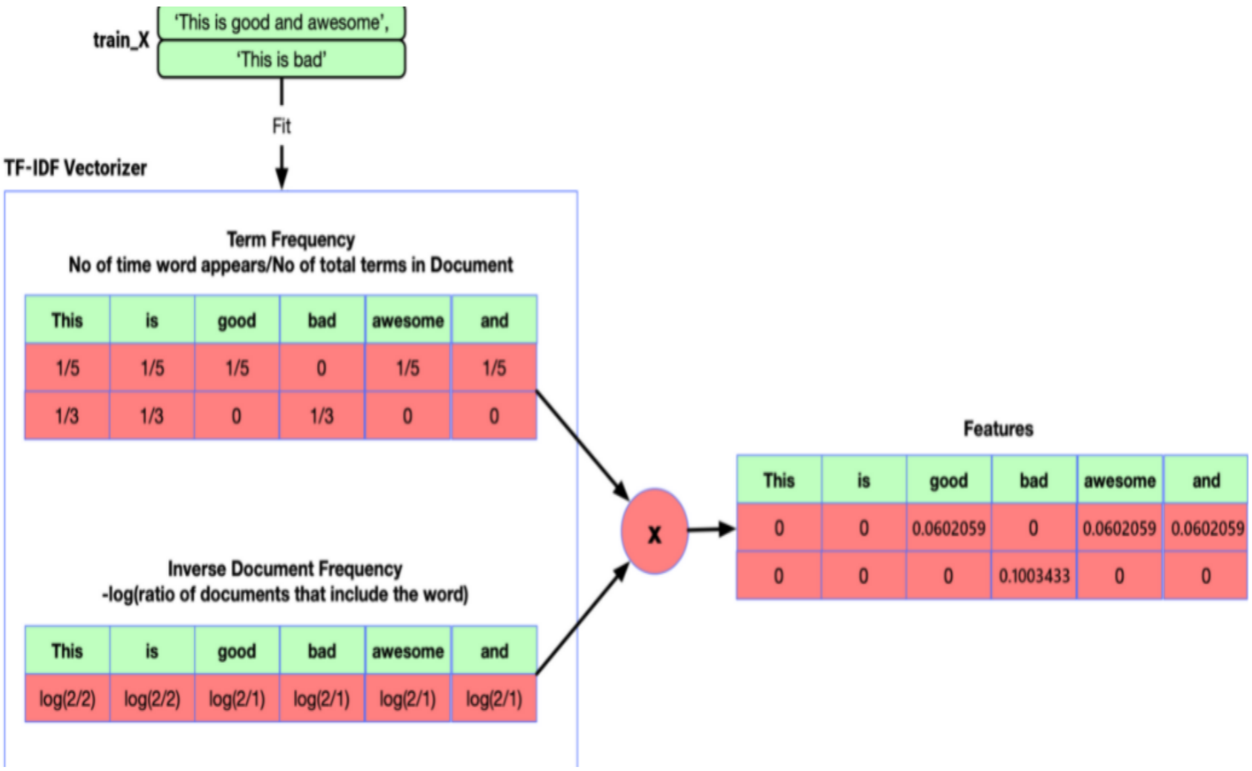
  Uppercase and lowercase letters could indeed coexist in phrases. A textual data is made up of various sentences. A most common method for reducing the complex problem would be to reduce everything to lowercase letters. This aligns all phrases in a document, but still it frequently modifications the meaning of the words, such as "US" to "us," in which the first signifies the United States of America and the other one is a pronoun.

- **TF-IDF Features :**

  The second method broadens the bag-of-words template by accounting for which included of words inside the corpuses. In this step, we convert our text to numeric values such that the Ai technologies could indeed handle it. Furthermore, the above method aids in penalizing overly commonly used words and providing more feature set. The following are the benefits of using this feature extractor:

  - Simple to compute.
  - It makes calculating the similarity of two documents simple.

- The most descriptive terms in a document are extracted using a basic metric.
- Because of IDF am, is, and so on, common words have no effect on the results.



- **Implementation :**
  - **Algorithm :**

    Step 1 : Reading our Data
    Step 2 : Data Preprocessing and EDA
    Step 3 : Splitting the data to Train and Test
    Step 4 : Implementing our models of summarization and categorization
    Step 5 : Training our models with the training data.
    Step 6 : Predicting our input data category and summarization

    Extra Steps for UI :
    Step 7 : Connecting our back end to the front end
    Step 8 : In our front end, we gave flexibility to the user to select their model for categorization and summarization.

- **Explanation of implementation :**

  - **Reading our Data :**

	Text	Category
0	Monsanto fined $1.5m for bribery\n\nThe US agr...	business
1	GM, Ford cut output as sales fall\n\nUS car fi...	business
2	Wal-Mart fights back at accusers\n\nTwo big US...	business
3	Rank 'set to sell off film unit'\n\nLeisure gr...	business
4	Georgia plans hidden asset pardon\n\nGeorgia i...	business
...	...	...
2219	Bond game fails to shake or stir\n\nFor gaming...	tech
2220	Europe backs digital TV lifestyle\n\nHow peopl...	tech
2221	The future in your pocket\n\nIf you are a geek...	tech
2222	California sets fines for spyware\n\nThe maker...	tech
2223	US duo in first spam conviction\n\nA brother a...	tech

2224 rows × 2 columns

  - **Data Preprocessing and EDA :**

	Text	Category	Processed Text
0	Monsanto fined $1.5m for bribery\n\nThe US agr...	business	monsanto fined bribery the us agrochemical gia...
1	GM, Ford cut output as sales fall\n\nUS car fi...	business	gm ford cut output sale fall us car firm gener...
2	Wal-Mart fights back at accusers\n\nTwo big US...	business	wal mart fight back accuser two big us name la...
3	Rank 'set to sell off film unit'\n\nLeisure gr...	business	rank set sell film unit leisure group rank cou...
4	Georgia plans hidden asset pardon\n\nGeorgia i...	business	georgia plan hidden asset pardon georgia offer...
...	...	...	...
2219	Bond game fails to shake or stir\n\nFor gaming...	tech	bond game fails shake stir for gaming fan word...
2220	Europe backs digital TV lifestyle\n\nHow peopl...	tech	europe back digital tv lifestyle how people re...
2221	The future in your pocket\n\nIf you are a geek...	tech	the future pocket if geek gadget fan next mont...
2222	California sets fines for spyware\n\nThe maker...	tech	california set fine spyware the maker computer...
2223	US duo in first spam conviction\n\nA brother a...	tech	us duo first spam conviction a brother sister ...

2224 rows × 3 columns

- **Splitting the data to Train and Test :**

```
Train size: 1779
Test size: 445
```

- **Implementing our models of summarization and categorization :**

```
Epoch 1/2
51/51 [==============================] - 15s 279ms/step - loss: 0.2741 - accuracy: 0.9382 - val_loss: 0.0789 - val_accuracy: 0.9719
Epoch 2/2
51/51 [==============================] - 14s 274ms/step - loss: 9.4453e-04 - accuracy: 1.0000 - val_loss: 0.0936 - val_accuracy: 0.9719
14/14 [==============================] - 1s 86ms/step - loss: 0.0934 - accuracy: 0.9685
Accuracy on test: 0.968539297580719
INFO:tensorflow:Assets written to: ram://4ed52282-0966-43e3-88eb-175a9844f619/assets
```

- **Predicting our input data category and summarization:**

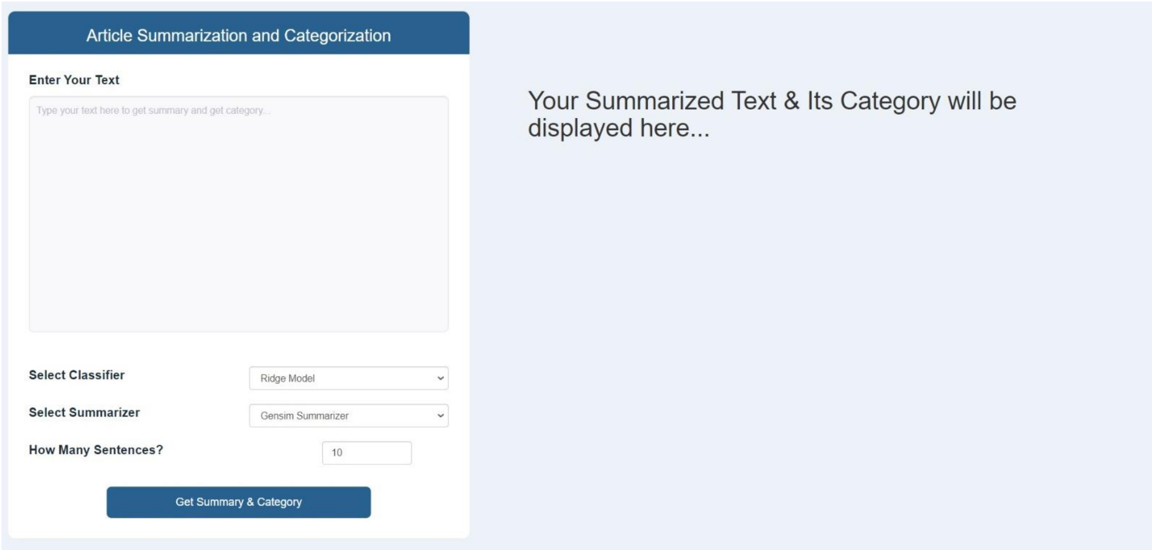
```
summerize_category(article_1, 10, ridge_model, "nltk")
```

```
----------------------------------------------------------------------------------------------------------------
Text category: Entertainment
----------------------------------------------------------------------------------------------------------------
Text summary :
After Washington's second Inaugural address, the next shortest was Franklin D. Roosevelt's fourth address on January 20, 1945, at just 559 words. John Adams' Inaugural address, which totaled 2,308 words, contained the longest sentence, at 737 words. Jackson became the first President to take his oath of office and deliver his address on the East Front Portico of the U.S. Capitol in 1829. The custom of delivering an address on Inauguration Day started with the very first Inauguration—George Washington's—on April 30, 1789. In 1921, Warren G. Harding became the first President to take his oath and deliver his Inaugural address through loud speakers. In 1925, Calvin Coolidge's Inaugural address was the first to be broadcast nationally by radio. There, Washington gave the shortest Inaugural address on record—just 135 words —before repeating the oath of office. With few exceptions, the next 37 Inaugurations took place there, until 1981, when Ronald Reagan's Swearing-In Ceremony and Inaugural address occurred on the West Front Terrace of the Capitol. William Henry Harrison delivered the longest Inaugural address, at 8,445 words, on March 4, 1841—a bitterly cold, wet day. And in 1949, Harry S. Truman became the first President to deliver his Inaugural address over television airwaves.
----------------------------------------------------------------------------------------------------------------
```

```
summerize_category(article_2, 5, our_model, "nltk", our=True)
```

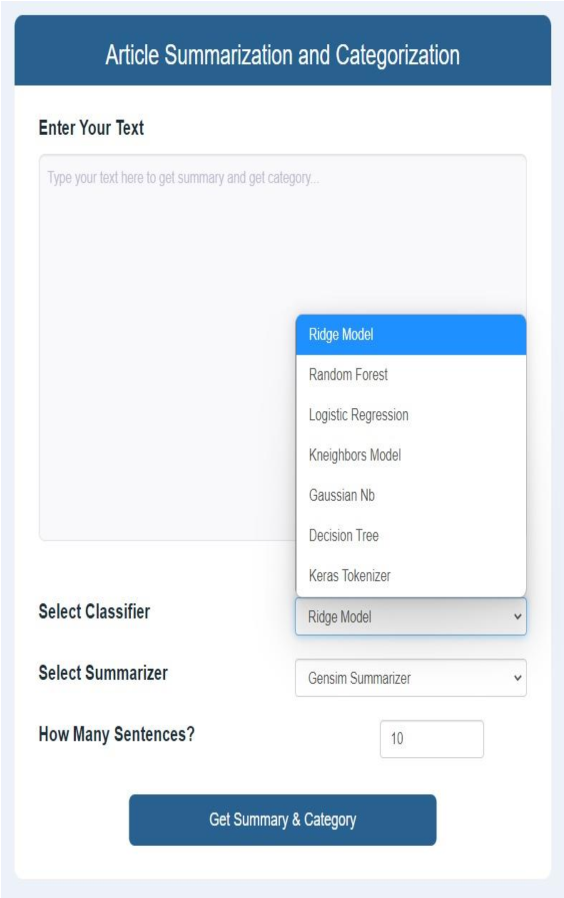
```
----------------------------------------------------------------------------------------------------------------
Text category: Politics
----------------------------------------------------------------------------------------------------------------
Text summary :
"I think we're going to get our amendment at a 50-vote threshold and what the vote count is going to be, I think we're still counting votes," he said. "I think that's why everyone's sitting around and standing around here still." "I think it's this evening," he added, when asked if he thinks the Senate will vote tonight on the CR. "The only thing I want to shut down is enforcement of an immoral, unconstitutional vaccine mandate," Lee said in Senate floor remarks. We're not going to shut the government down," the top Republican said in an interview on Fox News, adding "That makes no sense for anyone.
----------------------------------------------------------------------------------------------------------------
```

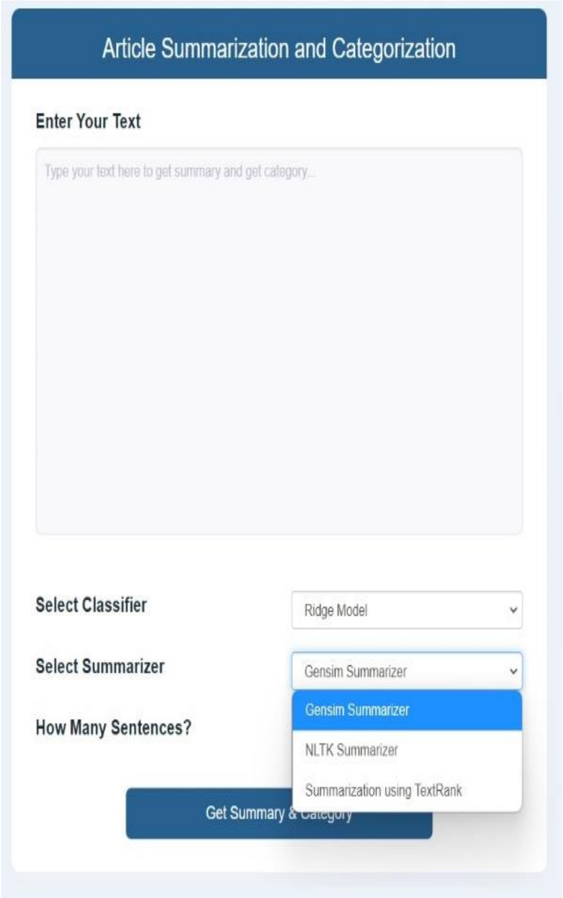
- **Connecting our back end to the front end :**
  - **Homepage of the application:**



  - **Our Models in UI :**

User can select any model of summarization and categorization.





- **Results :**

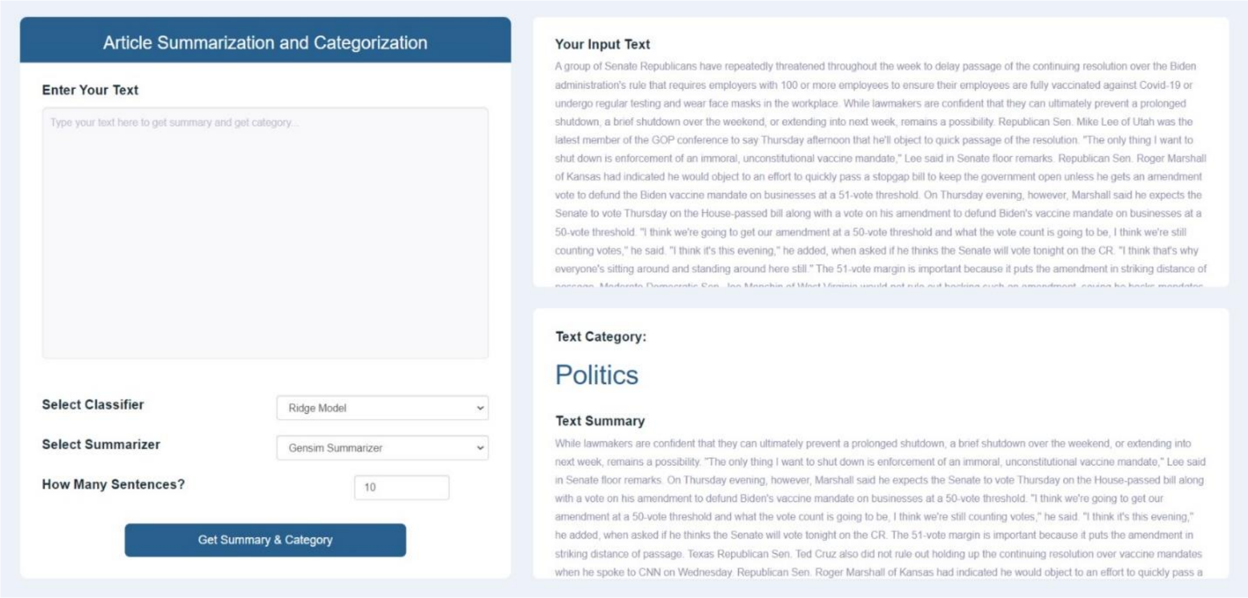
  - For our function summarize_category, we have given our summarization model as TextRank and Categorization model as Ridge. Our output is

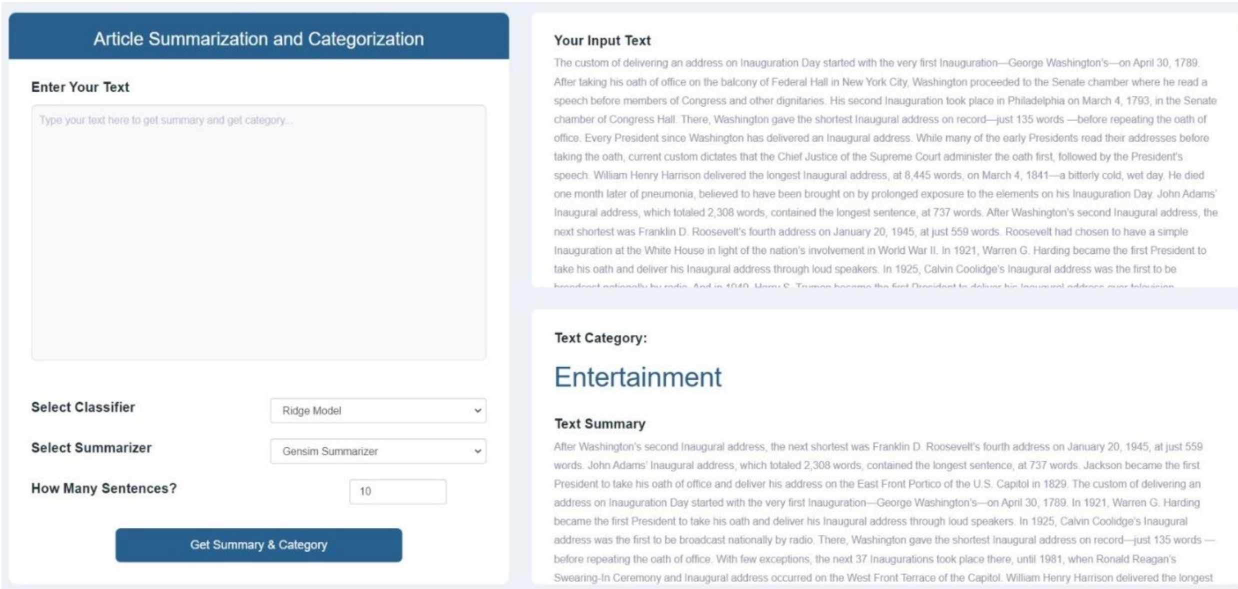
```
summerize_category(article_1, 10, ridge_model, "textrank")

-------------------------------------------------------------------------------------------------------------------------------
/usr/local/lib/python3.7/dist-packages/numpy/core/_methods.py:47: RuntimeWarning: overflow encountered in reduce
  return umr_sum(a, axis, dtype, out, keepdims, initial, where)
Text category: Entertainment
-------------------------------------------------------------------------------------------------------------------------------
Text summary :
The custom of delivering an address on Inauguration Day started with the very first Inauguration—George Washington's—on April 30, 1789. After taking his oath of
office on the balcony of Federal Hall in New York City, Washington proceeded to the Senate chamber where he read a speech before members of Congress and other
dignitaries. His second Inauguration took place in Philadelphia on March 4, 1793, in the Senate chamber of Congress Hall. There, Washington gave the shortest
Inaugural address on record—just 135 words —before repeating the oath of office. Every President since Washington has delivered an Inaugural address. While many of
the early Presidents read their addresses before taking the oath, current custom dictates that the Chief Justice of the Supreme Court administer the oath first,
followed by the President's speech. William Henry Harrison delivered the longest Inaugural address, at 8,445 words, on March 4, 1841—a bitterly cold, wet day. He died
one month later of pneumonia, believed to have been brought on by prolonged exposure to the elements on his Inauguration Day. John Adams' Inaugural address, which
totaled 2,308 words, contained the longest sentence, at 737 words. After Washington's second Inaugural address, the next shortest was Franklin D. Roosevelt's fourth
address on January 20, 1945, at just 559 words.
-------------------------------------------------------------------------------------------------------------------------------
```

  - For our function summarize_category, we have given our summarization model as Keras Tokenizer and Categorization model as Gensim. Our output is

```
summerize_category(article_2, 5, our_model, "gensim", our=True)

-------------------------------------------------------------------------------------------------------------------------------
Text category: Politics
-------------------------------------------------------------------------------------------------------------------------------
Text summary :
A group of Senate Republicans have repeatedly threatened throughout the week to delay passage of the continuing resolution over the Biden administration's rule that
requires employers with 100 or more employees to ensure their employees are fully vaccinated against Covid-19 or undergo regular testing and wear face masks in the
workplace.
Republican Sen. Roger Marshall of Kansas had indicated he would object to an effort to quickly pass a stopgap bill to keep the government open unless he gets an
amendment vote to defund the Biden vaccine mandate on businesses at a 51-vote threshold.
On Thursday evening, however, Marshall said he expects the Senate to vote Thursday on the House-passed bill along with a vote on his amendment to defund Biden's
vaccine mandate on businesses at a 50-vote threshold.
"I think that's why everyone's sitting around and standing around here still."
Texas Republican Sen. Ted Cruz also did not rule out holding up the continuing resolution over vaccine mandates when he spoke to CNN on Wednesday.
-------------------------------------------------------------------------------------------------------------------------------
```

UI Results :





<u>Project Management :</u>

- **Implementation Status Report :**
  - **Work Completed :**

    - **Description :** We have completed our task of categorization and summarization. We included more models for both summarization and categorization.

    - **Responsibility :**
      - Text summarization : Jaya Prakash Reddy, Divya Geethanjali
      - Text Classification and training the model : Divya Geethanjali and Ayyappa
      - UI : Jaya Prakash, Ayyappa
      - Documentation : Divya Geethanjali, Jaya Prakash and Ayyappa.

    - **Contribution :**
      - Divya Geethanjali Birudharaju : 33%
      - Jaya Prakash Reddy Gade : 33.5%
      - Ayyappa Ankishetty : 33.5%

    - **Issues/Concerns :**
      - As we had many models, it was hard for us to understand about each model and implement it in our code.
      - Each model has issues in running in Colab, we had to search tutorials and learn to avoid those issues
      - As each one of us were busy with their schedule, we had to complete the works on each individual flexible time and update each other.
      - As we also wanted to implement UI, it was hard to copy codes and segregate them to each model and connect it to front end.

Github : https://github.com/jayaprakashreddy007/NLP/tree/main/Project

Video Link :

https://unt.zoom.us/rec/share/fJ08Y4_TXx5EXCG3q-lp-zbD6nw7kU7dQGQlyPZ7EX4vQCvtBQJqwoPj4fq6MOv8.59tTGOknILLF860u?startTime=1638509669000

References :

- M.A. Mosa, A.S. Anwar, A. Hamouda, A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms. Knowl.-Based Syst. 163, 518–532 (2019)
- R. Barzilay, M. Elhadad, Using lexical chains for text summarization. Adv. Autom. Text Summ. 111–121 (1999)
- F.B. Goularte, S.M. Nassar, R. Fileto, H. Saggion, A text summarization method based on fuzzy rules and applicable to automated assessment. Expert Syst. Appl. 115, 264–275 (2019)
- S. Shashikanth, S. Sanghavi, Text summarization techniques survey on telugu and foreign languages (2019)