**(Times Internet Limited)  
 Campus Live Project**

IAB CATEGORIES REQUIRED:

Academic Interests,Automotive ,Books and Literature,Business and Finance,Careers,Family and Relationships,Fine Art,Food & Drink,Health and Medical Services ,Healthy Living,Hobbies & Interests,Home & Garden,Medical Health,Movies,Music and Audio,News and Politics,Personal Finance ,Pets,Pharmaceuticals, Conditions, and Symptoms,Pop Culture,Real Estate,Shopping,Sports,Style & Fashion,Technology & Computing,Television,Travel,Video Gaming

Dataset 1 (Hugging Face) <https://huggingface.co/datasets/kinnews_kirnews/tree/main>

Downloaded from-><https://github.com/Andrews2017/KINNEWS-and-KIRNEWS-Corpus>

Description of the Dataset

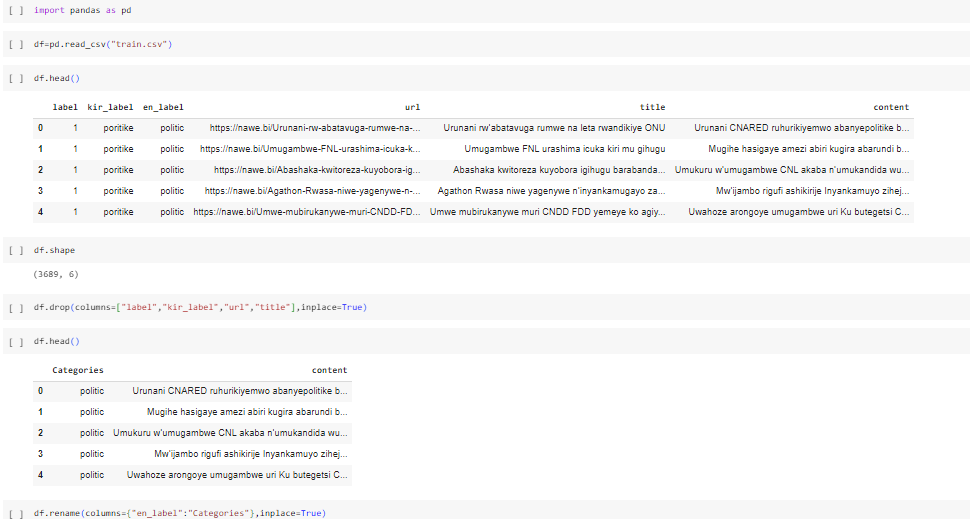
This dataset can be used for text classification of news articles in Kinyarwadi and Kirundi languages. Each news article can be classified into one of the 14 possible classes. The classes are: politics,sport,economy,health,entertainment,history,technology,culture,religion,environment,education,relationship.

* Firstly I translate the whole dataset from Kinyarwadi and Kirundi languages to English Languages through using Google Translate.
* Then I preprocess the dataset through converting into dataframe and perform the cleaning steps i.e droping the irrelevant columns,filling the null values,etc.
* Now,finally I make two column i.e Text and Category in the Dataframe.

Text corresponds to these categories.

Finally Convert the dataframe into csv file.

**Implementation**

l.j



* Google colab-> <https://colab.research.google.com/drive/1FNwLaC-lueiT-Mo1BzW6-r3fVSM9Bii_#scrollTo=VNJB8sdDTDmc>
* <https://colab.research.google.com/drive/1ZMnnL7LSlRbcZesP752JAdvz0UtVpvPZ>
* Total Unique IAB Categories that we required from this is-> 6 categories or classes[health,politics,sports,technology,culture,education]

**Dataset 2(Kaggle)**

<https://www.kaggle.com/competitions/learn-ai-bbc/data?select=BBC+News+Train.csv>

Description of the Dataset(BBC News):

* Data fields
* ArticleId - Article id unique # given to the record
* Article - text of the header and article
* Category - cateogry of the article (tech, business, sport, entertainment, politics/li>
* 5 Unique classes

Steps->

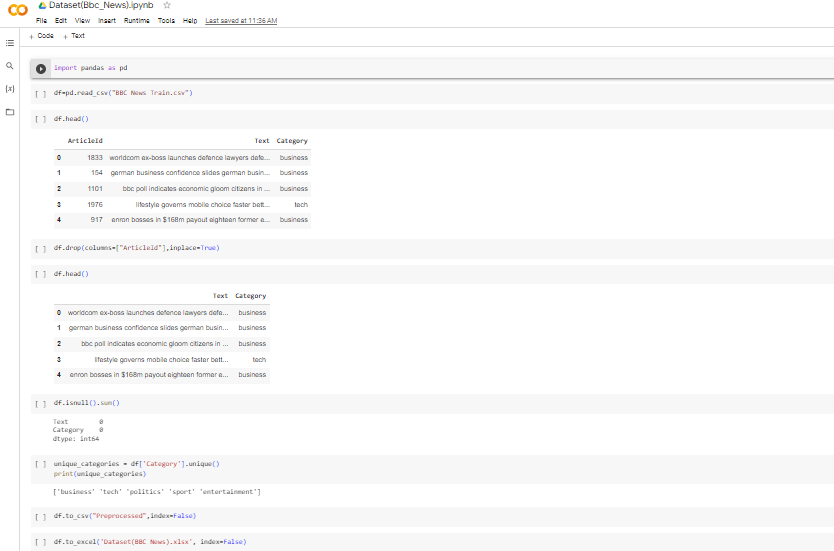
* Firstly I collect the dataset from kaggle .It is Bbc News Category Dataset
* Then I preprocess the dataset through converting into dataframe and perform the cleaning steps i.e droping the irrelevant columns,filling the null values, and some other steps etc.
* Now,finally I make two column i.e Text and Category in the Dataframe.

Text corresponds to these categories.

Finally Convert the dataframe into csv file.

* Google Colab-> <https://colab.research.google.com/drive/1jIHxb6OD70aVb5e8LTXbOWYVGtt7Mzgt>

**Implementation**



**Dataset 3(Kaggle)**

<https://www.kaggle.com/code/yufengdev/bbc-text-categorization/input>

Description of the Dataset(BBC News2):

* Data fields
* ArticleId - Article id unique # given to the record
* Article - text of the header and article
* Category - cateogry of the article (tech, business, sport, entertainment, politics/li>
* 5 Unique classes

Steps->

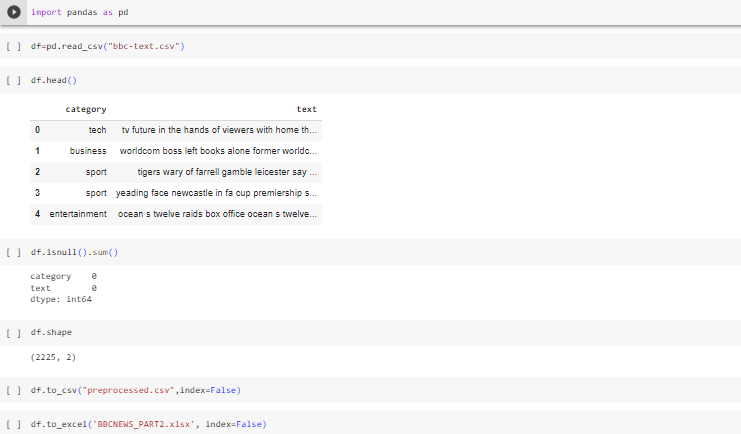
* Firstly I collect the dataset from kaggle .It is Bbc News Category Dataset
* Then I preprocess the dataset through converting into dataframe and perform the cleaning steps i.e droping the irrelevant columns,filling the null values, and some other steps etc.
* Now,finally I make two column i.e Text and Category in the Dataframe.

Text corresponds to these categories.

Finally Convert the dataframe into csv file.

* Google Colab-> <https://colab.research.google.com/drive/1DUYTInfVPWupRrvJwgwp3g0R5vWJOs4z#scrollTo=-jOffgkXshbN>

**Implementation**



Dataset 4 (Kaggle)

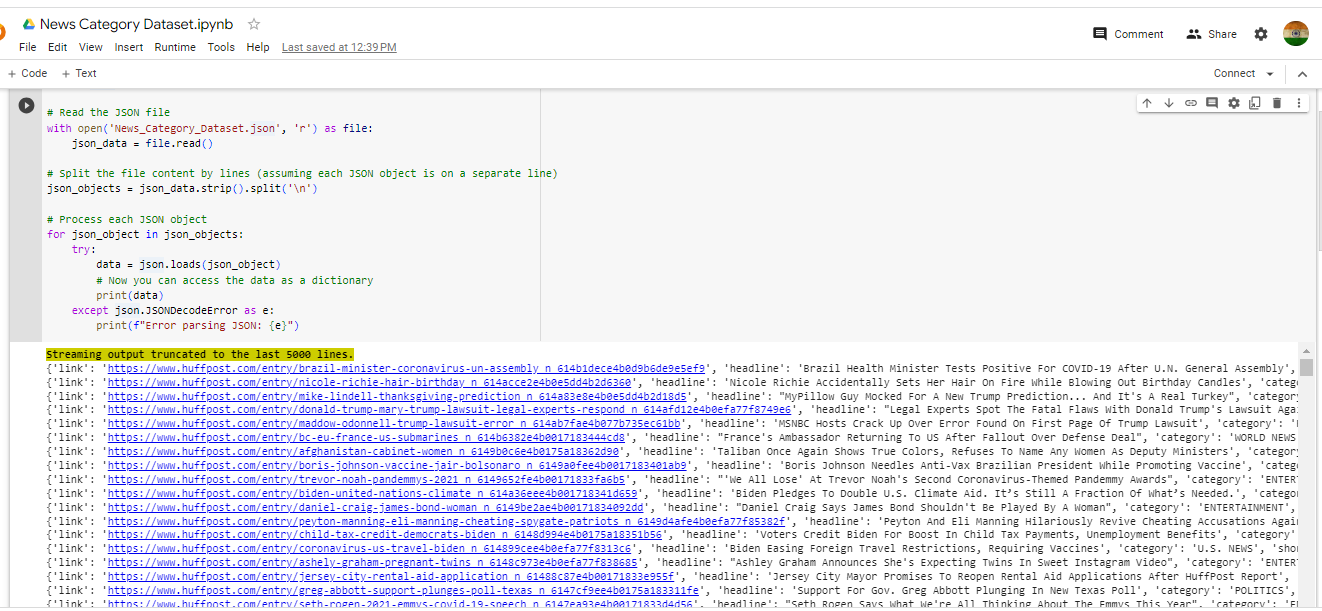
<https://www.kaggle.com/datasets/rmisra/news-category-dataset>

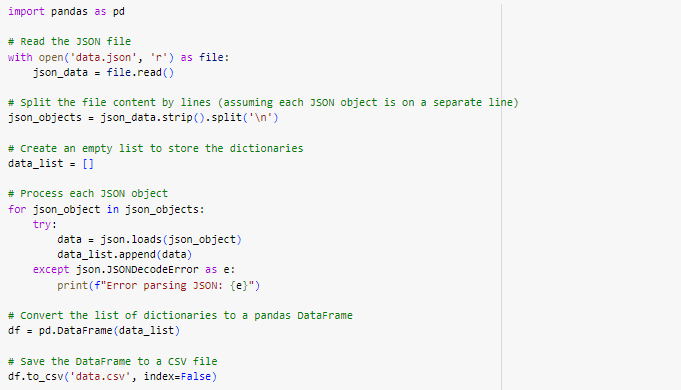
Description of the Dataset(NEWS CATEGORY DATASET):

* Each record in the dataset consists of the following attributes:
* category: category in which the article was published.
* headline: the headline of the news article.
* authors: list of authors who contributed to the article.
* link: link to the original news article.
* short\_description: Abstract of the news article.
* date: publication date of the article.
* There are a total of 42 news categories in the dataset. The top-15 categories and corresponding article counts are as follows:
* POLITICS: 35602
* WELLNESS: 17945
* ENTERTAINMENT: 17362
* TRAVEL: 9900
* STYLE & BEAUTY: 9814
* PARENTING: 8791
* HEALTHY LIVING: 6694
* QUEER VOICES: 6347
* FOOD & DRINK: 6340
* BUSINESS: 5992
* COMEDY: 5400
* SPORTS: 5077
* BLACK VOICES: 4583
* HOME & LIVING: 4320
* PARENTS: 3955

The Dataset is a json file and first I convert into csv file.

**Implementation**





Steps->

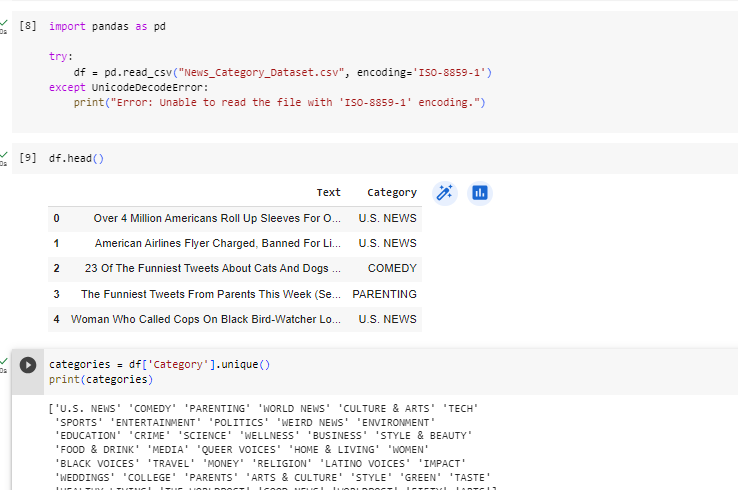
* Firstly I collect the dataset from kaggle .It is Bbc News Category Dataset
* Then I preprocess the dataset through converting into dataframe and perform the cleaning steps i.e drop the irrelevant columns like url,title etc ,filling the null values, and some other steps etc.
* Now,finally I make two column i.e Text and Category in the Dataframe.

Text corresponds to these categories.

Finally Convert the dataframe into csv file.

**Google Colab**-> <https://colab.research.google.com/drive/1PYuHG-EwLr0ltUUAhaiJ6yG4RHcA83Xs>

Total Unique Categories->41

* ['U.S. NEWS' 'COMEDY' 'PARENTING' 'WORLD NEWS' 'CULTURE & ARTS' 'TECH' 'SPORTS' 'ENTERTAINMENT' 'POLITICS' 'WEIRD NEWS' 'ENVIRONMENT' 'EDUCATION' 'CRIME' 'SCIENCE' 'WELLNESS' 'BUSINESS' 'STYLE & BEAUTY' 'FOOD & DRINK' 'MEDIA' 'QUEER VOICES' 'HOME & LIVING' 'WOMEN' 'BLACK VOICES' 'TRAVEL' 'MONEY' 'RELIGION' 'LATINO VOICES' 'IMPACT' 'WEDDINGS' 'COLLEGE' 'PARENTS' 'ARTS & CULTURE' 'STYLE' 'GREEN' 'TASTE''HEALTHY LIVING' 'THE WORLDPOST' 'GOOD NEWS' 'WORLDPOST' 'FIFTY' 'ARTS']
* Total Unique IAB Categories that we required from this is-> 14 categories
* 

For further processing I explored some of the model-> <https://www.kaggle.com/code/nicoladisabato/roberta-bert-xlnet-0-73-accuracy>

**Merged Dataset(Final)**

* Merged all the above dataset and convert it into Merged final Dataset
* Doing some preprocessing steps over it and finally obtained required columns i.e.Text and Categories

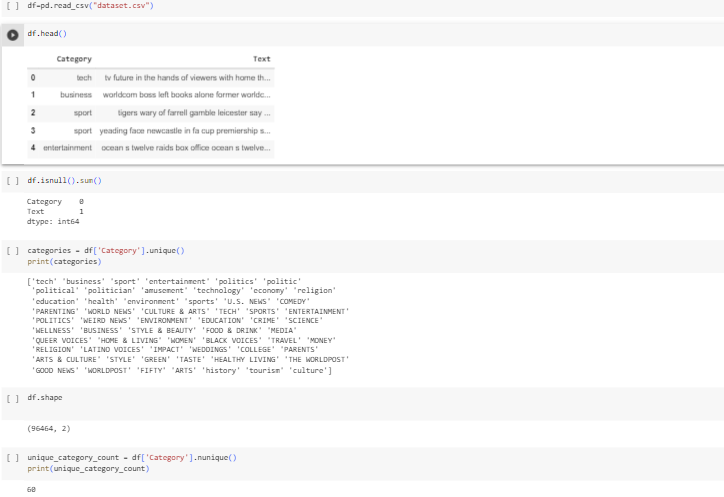
Total Unique Categories->60

* ['tech' 'business' 'sport' 'entertainment' 'politics' 'politic'
* 'political' 'politician' 'amusement' 'technology' 'economy' 'religion'
* 'education' 'health' 'environment' 'sports' 'U.S. NEWS' 'COMEDY'
* 'PARENTING' 'WORLD NEWS' 'CULTURE & ARTS' 'TECH' 'SPORTS' 'ENTERTAINMENT'
* 'POLITICS' 'WEIRD NEWS' 'ENVIRONMENT' 'EDUCATION' 'CRIME' 'SCIENCE'
* 'WELLNESS' 'BUSINESS' 'STYLE & BEAUTY' 'FOOD & DRINK' 'MEDIA'
* 'QUEER VOICES' 'HOME & LIVING' 'WOMEN' 'BLACK VOICES' 'TRAVEL' 'MONEY'
* 'RELIGION' 'LATINO VOICES' 'IMPACT' 'WEDDINGS' 'COLLEGE' 'PARENTS'
* 'ARTS & CULTURE' 'STYLE' 'GREEN' 'TASTE' 'HEALTHY LIVING' 'THE WORLDPOST'
* 'GOOD NEWS' 'WORLDPOST' 'FIFTY' 'ARTS' 'history' 'tourism' 'culture']
* Total Rows are 96464 and columns are 2

Total Unique IAB Categories that we required from this is-> 18 categories

**Implementation**





**Google Colab**-> <https://colab.research.google.com/drive/1GP4hDxzwO2RiyDUO4RLcExVDTo0v-De8#scrollTo=ylgysJpvGIpB>

**Aman Rai –**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **IAB Category** | Dataset 1 (Hugging Face) | **Dataset 2 (Kaggle)** | **Dataset 3**  **(Kaggle)** | **Merged Dataset** |
| Academic Interests | History, Economy, Education |  | Education, Science, College | Economy,Education,College,History,Science |
| Automotive |  |  |  |  |
| Books and Literature |  |  |  |  |
| Business and Finance | Business | Business | Business | Buisness |
| Careers |  |  |  |  |
| Family and Relationships |  |  | Parenting, Parents | Parenting,Parents,Women |
| Fine Art |  |  | Arts | Arts |
| Food & Drink |  |  | Food & Drink,Taste | Food & Drink,Taste |
| Health and Medical Services | Health |  |  | Health |
| Healthy Living |  |  | Healthy Living, Wellness | Healthy Living,Wellness |
| Hobbies & Interests |  |  |  |  |
| Home & Garden |  |  | Homes & Living,  Environment | Homes &Living,  Environment |
| Medical Health |  |  |  |  |
| Movies | entertainment | entertainment | Entertainment, Comedy | Entertainment, Comedy |
| Music & Audio |  |  | Queer Voices, Black Voices, Latino Voices | Queer Voices, Black Voices, Latino Voices |
| News & Politics | Politics | Politics | Politics, US News, World News, Weird News, Good News | Politics, Politic,Political,Politician,US News, World News, Weird News, Good News |
| Personal Finance |  |  | Money | Money |
| Pets |  |  |  |  |
| Pharmaceuticals, Condition & Symptoms |  |  |  |  |
| Pop Culture | Culture |  | Arts and Culture | Art & Culture |
| Real Estate |  |  |  |  |
| Shopping |  |  |  |  |
| Sports | Sport | Sport | Sports | Sports,sport |
| Style & Fashion |  |  | Style & Beauty, Style | Style & Beauty,Style |
| Technology & Computing | Technology |  | Tech | Tech,Technology |
| Television |  |  | Media | Media |
| Travel | Tourism |  | Travel | Travel,Tourism |
| Video Gaming |  |  |  |  |

We have merged the category Fine Art and Pop Culture into Arts and Culture & Health and Medical Services & Medical Health  into Health.

**Our Objectives**

1-Problem of Imbalances datasets->Solved by data Augumentation techniques

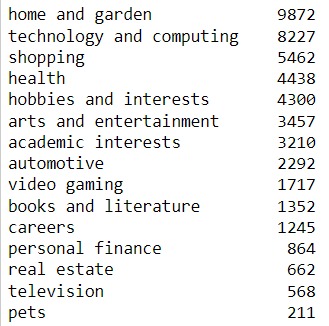
2-Count the Unique Classess in fetched Datasets

3-Explore Some Preprocessing Techniques

We get total Counts of 26 Category-



**Challenges->**These Categories Have less counts so we have to focus on this



We could find the way through->

Certainly, here are potential sources and ideas for datasets related to the topics "Careers," "Pets," "Real Estate," "Books and Literature," "Video Gaming," and "Automotive":

Careers: <https://data.world/jobspikr/10000-data-scientist-job-postings-from-the-usa>

Books and Literature:

<https://www.kaggle.com/datasets/thedevastator/comprehensive-literary-greats-dataset>

VideoGaming:

Automotive:

<https://huggingface.co/datasets/SerhiiBond/automotive_sentiment_analysis>

2-Pets Datasets Have images only

3-Not a particular datasets of television found

4-In video gaming Category->We have either found numerical datasets or dialogue naration of characters in a particular datasets.

<https://www.kaggle.com/datasets/atharvaingle/video-games-dataset>

<https://github.com/YipingNUS/nle-supplementary-dataset>

5- Resources(GPU)->System is slow while working on large datasets.

**Text Classification of News Articles**

<https://www.analyticsvidhya.com/blog/2021/12/text-classification-of-news-articles/>

**Text Classification**

Text classification datasets are used to categorize natural language texts according to content. Forexample, think classifying news articles by topic, or classifying book reviews based on a positive or negative response. Text classification is also helpful for language detection, organizing customer feedback, and fraud detection.While this process is time-consuming when done manually, it can be automated with machine learning models.Category classification, for news, is a multi-label text classification problem. The goal is to assign one or more categories to a news article. A standard technique in multi-label text classification is to use a set of binary classifiers.

**Data Fields**

* Article Id – Article id unique given to the record
* Article – Text of the header and article
* Category – Category of the article (tech, business, sport, entertainment, politics)

**Data Cleaning and Data Preprocessing**

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

* **Import Libraries**

1. let’s import the necessary Python libraries and the dataset that we need for this task.

* **Import Dataset**

1. Import the dataset which we will have to use. (BBCNews.csv)

* **Shape of Dataset**

1. Check the shape (row and column) of the dataset

1. Check Information of Columns of Dataset
2. Columns of Dataset
3. Count Values of Categories

* **Convert Categories Name into Numerical Index**

1. Convert the given news categories into categorical values.Associate Category names with numerical index and save it in new column CategoryId

* **Show Category’s Name w.r.t Category ID**

1. Here you can show that news category’s name with respect to the following unique category ID. Create a new pandas dataframe "category", which only has unique Categories, also sorting this list in order of CategoryId values

**Exploratory Data Analysis (EDA**)

1. In data mining, Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modeling task.

* **Visualizing Data**

1. The below graph shows the news article count for category from our dataset.

Show Text Column of Dataset

* **Tokenization**

1. In natural language processing, tokenization is the text preprocessing task of breaking up text into smaller components of text (known as tokens).

* **Remove All Tags**

1. First, we remove all tags which are present in our given dataset

* **Remove Special Characters**

1. Here we remove all the special characters

* **Convert Everything in Lower Case**

1. We convert all articles or text to lower case. It is one of the simplest and most effective forms of text preprocessing. It is applicable to most text mining and NLP problems and can help in cases where your dataset is not very large and significantly helps with the consistency of expected output

* **Remove all Stopwords**

1. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words to take up space in our database, or take up the valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages.

* **Lemmatizing the Words**:

1. Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So, it links words with similar meanings to one word. lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

* **Part-of-Speech Tagging:**

1. In natural language processing, part-of-speech tagging is the process of assigning a part of speech to every word in a string. Using the part of speech can improve the results of lemmatization.

* **Declared Dependent and Independent Value**
* **Create and Fit Bag of Words Model**

1. In this step, we construct a vector, which would tell us whether a word in each sentence is a frequent word or not. If a word in a sentence is a frequent word, we set it as 1, else we set it as 0. Whenever we apply any algorithm in NLP, it works on numbers. We cannot directly feed our text into that algorithm. Hence, the Bag of Words model is used to preprocess the text by converting it into a bag of words, which keeps a count of the total occurrences of the most frequently used words.

* **Train Test and Split the Dataset**

1. We need to split a dataset into train and test sets to evaluate how well our machine learning model performs. The train set is used to fit the model, the statistics of the train set are known. The second set is called the test data set, this set is solely used for predictions.
2. Create, Fit and Predict all ML Model
3. Create Dataframe of Model, Accuracy, Precision, Recall, and F1
4. Best Model to Perform Accuracy Score
5. Fit & predict best ML Model
6. Predict News Article : Now, here, after the completion of model analysis, we can also predict any news articles.

**Conclusion**

Finally after doing Data cleaning and Data Preprocessing (cleaning data, train\_test\_split model, creating a bag of words NLP model, and machine learning model) we got the accuracy scores and we can say that Random Forest Classification(In this particular article i.e BBC NEWS.csv dataset) gives the best accuracy among all machine learning models.

We have set the count of each IAB category data to 20,000. I have scrapped data from these websites like Times of India,Economic Times etc**.**

**Web-Scraping**

Web scraping is a technique used in computer programming to extract information and data from websites. It involves automated processes that retrieve specific content, such as text, images, links, and other structured data, from web pages. This data can then be analyzed, processed, and utilized for various purposes, such as data aggregation, research, analysis, content extraction, or populating databases.

**Following Methodologies have been employed in order to extract Raw-Datasets using Web-Scraping:**

1. Employed the website's sitemap file to methodically retrieve URLs, harnessing its organized structure designed to optimize both search engine indexing efficiency and user-friendly navigation. The extracted URLs were systematically exported to a CSV file for further analysis.
2. Utilized the Beautiful Soup library to proficiently extract a comprehensive list of URLs. This library's capabilities were harnessed to navigate through the web page's structure, enabling the accurate and efficient compilation of URLs for subsequent actions.
3. Leveraged the advanced features of the Beautiful Soup library to meticulously extract textual content from the website's URLs, ensuring a precise extraction of article text grouped by specific categories. This approach facilitated the creation of a topic-wise corpus, conducive to refined content analysis and exploration.

So we have to applied various Data-Augumentation Techniques to increase the count of these data of a particular category in the dataset.

**Data Augumentation Techniques**

1. **Synonym replacement**

Synonym replacement is a data augmentation technique used to enhance text data for natural language processing tasks. In this technique, certain words within the text are replaced with their synonyms, maintaining the contextual meaning while introducing variability in the dataset. This helps improve the robustness and generalization of machine learning models by exposing them to different word choices that convey similar meanings. The process involves identifying synonyms through linguistic resources or embeddings, and selectively substituting words while preserving the overall structure and semantics of the text.

1. **Random Deletion**

Random deletion is a data augmentation technique employed to diversify text data for natural language processing tasks. In this approach, words or tokens within a sentence are randomly removed with a certain probability. By selectively eliminating portions of the text, the technique introduces noise and encourages the model to rely on contextual cues and neighboring words for comprehension. This process aims to enhance the model's ability to handle missing information and improve its overall robustness by training it on a broader range of potential input variations.

1. **Random Insertion**

Random insertion is a data augmentation technique used to enrich text data for natural language processing tasks. In this method, new words or tokens are randomly inserted into a sentence with a certain probability. By introducing additional content, the technique increases the diversity of the dataset and exposes the model to novel word combinations. This helps the model become more adaptable to unexpected inputs and enhances its ability to handle variations in text structure and vocabulary. Random insertion encourages the model to learn to incorporate new information while maintaining the context and semantics of the original text.

1. **Back Translation**

Back translation is a data augmentation technique commonly employed in natural language processing. It involves translating a piece of text from one language to another and then translating it back to the original language. This process generates a new version of the text that retains the original meaning but may have different phrasing and word choices. Back translation helps diversify the training data, enabling models to learn from various sentence structures and vocabulary

**Count of Categories-Data Obtained from Data-Augmentation Technique**

