**News classification**

So to there are several steps i.e.

**DATA ACQUISITION:-**

here the data is not given to us to and there are 28 categories under which we have to classify them so there are several things

1. Existing dataset:([code link](https://github.com/shishir-dwi/TIL-NLP-Project/blob/09ee824b057c9ca89f5b4627bb426ea6c37e119e/data%20aquisition%20-%20from%20preexisting%20data.ipynb))

From different websites like kaggle, UCI, etc

[News Aggregator Dataset | Kaggle](https://www.kaggle.com/datasets/uciml/news-aggregator-dataset) (4 Categories)

[News Category Dataset | Kaggle](https://www.kaggle.com/datasets/rmisra/news-category-dataset) (42 Categories)

[20 Newsgroups | Kaggle](https://www.kaggle.com/datasets/crawford/20-newsgroups) (20 categories)

[ag\_news · Datasets at Hugging Face](https://huggingface.co/datasets/ag_news/viewer/default/train?p=1199) (4 categories)

[N24News.zip - Google Drive](https://drive.google.com/file/d/1OS1fXwZ1Vsj70lEQajccyssxQRYp5X9D/view) (4 categories)

These are some dataset which don’t perfectly aligns with our categories but some the categories are usefull for us

Now data that we get is text and not labelled and now we have to categorize them according to our 28 categories

There are several methods to do that

1. Manual Annotation:

(what):- Manual annotation refers to the process of assigning labels or annotations to data manually by human annotators. In the context of machine learning and natural language processing (NLP), it involves human experts or workers reviewing data samples and assigning specific labels or categories to them.

1. Crowd Sourcing:

(what):-crowdsourcing refers to the process of obtaining human input or annotations from a large and diverse group of individuals, often referred to as the "crowd," to label or categorize the unlabelled news articles in your dataset. This approach leverages the collective efforts of multiple contributors to efficiently assign category labels to a large amount of data.

1. Data generation:



DAY 2:

1. Unique dataset

|  |  |  |
| --- | --- | --- |
| **IAB Category** | **Dataset 1** | **Dataset 2** |
| Academic Interests | Science and Education | ->EDUCATION  ->COLLEGE |
| Automotive | ->Auto & Vehicle  ->Vehicles |  |
| Books and Literature | Books & Literature |  |
| Business and Finance | ->Business and Industrial  ->Finance | BUSINESS |
| Careers | ->Job & Education  ->Jobs and Career |  |
| Family and Relationships |  | ->PARENTING  ->PARENTS  ->DIVORCE  ->WEDDING |
| Fine Art |  | ->ARTS & CULTURE |
| Food & Drink | ->Food & Drink  ->Food and Drink | ->FOOD & DRINKS  ->TASTE |
| Health and Medical Services | Health |  |
| Healthy Living | ->Beauty & Fitness  ->Lifestyle | ->WELLNESS  ->HEALTHY LIVING |
| Hobbies & Interests | ->Hobbies & Leisure  ->Hobbies and Leisure |  |
| Home & Garden | ->Home & Garden  ->Home and Garden | HOME & LIVING |
| Medical Health |  |  |
| Movies | Arts & Entertainment | ->ENTERTAINMENT  ->COMEDY |
| Music & Audio |  |  |
| News & Politics | ->Law & Government  ->News  -\Law and Government | ->POLITICS  ->THE WORLDPOST  ->CRIME  ->WORLD NEWS  ->WEIRD NEWS  ->WORLDPOST  ->GOOD NEWS  -> U.S. NEWS |
| Personal Finance |  |  |
| Pets | ->Pets & Animal  ->Pets and Animal |  |
| Pharmaceuticals, Condition & Symptoms |  |  |
| Pop Culture |  | CULTURE & ART |
| Real Estate | Real Estate |  |
| Shopping | ->Shopping  ->eCommerce & Shopping |  |
| Sports | Sports | SPORTS |
| Style & Fashion |  | ->STYLE & BEAUTY  ->STYLE |
| Technology & Computing | ->Computer Electronics  & Technology  ->Computer & Electronics | TECH |
| Television |  |  |
| Travel | Travel | TRAVEL |
| Video Gaming | Games |  |

To address the issue of having a limited dataset, we have adopted a technique known as Data Augmentation.

Data Augmentation ([code link](https://github.com/shishir-dwi/TIL-NLP-Project/blob/09ee824b057c9ca89f5b4627bb426ea6c37e119e/data%20aquisition%20-%20data%20augmentation.ipynb))

This method involves manipulating the existing data to create new instances. By applying transformations like rotation, scaling, and flipping to the original data, we generate a larger and more diverse dataset. This expanded dataset provides our models with a wider range of examples to learn from, which is particularly beneficial when working with constrained data. Data Augmentation not only aids in preventing overfitting but also enhances the model's ability to handle variations in real-world scenarios, ultimately leading to improved generalization and performance during training.

However, we found that Data Augmentation didn't yield the desired outcomes. Hence, we opted for an alternative approach. This involved scraping URLs from various websites, beginning with the Times of India website. By gathering data from diverse sources, we aimed to enhance the richness and comprehensiveness of our dataset, potentially overcoming the limitations posed by our initial data constraints.

Web scraping:([URL scraping code link](https://github.com/shishir-dwi/TIL-NLP-Project/blob/09ee824b057c9ca89f5b4627bb426ea6c37e119e/data%20aquisition%20-%20scraping%20article%20from%20url.ipynb)) ([Article scraping from URL code link](https://github.com/shishir-dwi/TIL-NLP-Project/blob/09ee824b057c9ca89f5b4627bb426ea6c37e119e/data%20aquisition%20-%20scraping%20url%20from%20website.ipynb))

In order to expand the dataset's scope, we implemented a web scraping initiative. By extracting URLs from prominent news websites, we systematically retrieved news articles' textual content using HTML parsing techniques. This enriched dataset infusion not only increased data volume but also introduced diverse perspectives from various reputable sources(like-Time of India, or Economics Times).

([URL scraping from different website’s sitemap code link](https://github.com/shishir-dwi/TIL-NLP-Project/blob/09ee824b057c9ca89f5b4627bb426ea6c37e119e/data%20aquisition%20-%20scraping%20url%20from%20website's%20xml%20file.ipynb))

To streamline our web scraping endeavor, we employed the websites' sitemap as a foundational resource. By retrieving the XML file of the news websites' sitemap, we obtained a structured outline of the website's content hierarchy. Leveraging this sitemap, we systematically extracted URLs corresponding to news articles. This approach allowed us to ensure comprehensive coverage of articles across categories and timelines, enhancing the representatives of the collected data for our news classification project

Using the above URL scraping techniques we have

|  |  |
| --- | --- |
| **CATEGORIES** | **NUMBER OF DATA SCRAPED** |
| PETS | 23468 |
| REAL ESTATE | 38278 |
| SHOPPING | 18811 |
| PERSONAL FINANCE | 14375 |
| TRAVEL | 13899 |
| VIDEO GAMING | 171 |
| TELEVISION | 23547 |