**Abstract:**

The primary objective of our project is to analyze news headlines and articles collected from various sources, from several nationalities, spanning across multiple languages to understand the coverage of various topics by different media sources in different countries. In order to do so, we are utilizing various publicly available datasets, cleaning them and merging them into one de-normalized dataset. In addition to this, we intend on scraping certain news websites in order to obtain the necessary information to complete our project. Upon collecting, claning, and condensing the dataset, we plan to use several machine learning and deep learning algorithms in order to obtain insights regarding the prominent topics covered (through topic modelling), obtain a summary of the article through text summarization, get an idea about the sentiments and emotions exhibited in each headline and article summary. Upon obtaining these metrics, we will combine them with our dataset and store it in 5 different formats: sqlite, csv, excel, json, and in the cloud using AWS(S3-for easy access). Finally, we plan to utilize classification algorithms to predict for an unknown news article, the news outlet and country of origin.

**1. Documenting data sources**

* Data Sources
* Describe the format of the raw data
* Describe the nature of the data (e.g., what it contains, how often it is updated, etc).

Datasets Used:

1. All the News Dataset: <https://components.one/datasets/all-the-news-2-news-articles-dataset>
   1. The "All the News 2.0" dataset contains 2.7 million news articles and essays from 27 American publications, dating from January 1, 2016, to April 2, 2020. The raw data is formatted with fields for date, year, month, day, author, title, article text, URL, section, and publication. The dataset, updated last on July 9, 2022, is available as a CSV file and does not indicate ongoing updates beyond this point
2. All the News 1.0: <https://components.one/datasets/all-the-news-articles-dataset>
   1. The "All the News 1.0" dataset features 204,135 articles from 18 American publications, primarily spanning from 2013 to early 2018. The data includes fields such as date, title, publication, article text, and URL, where available. It's formatted as a SQLite database, about 1.5 GB in size, and was last updated on January 19, 2019
3. Saudi News Net: <https://github.com/inparallel/SaudiNewsNet>
   1. The SaudiNewsNet dataset contains 31,030 Arabic newspaper articles with metadata from various Saudi newspapers. The data is stored in JSON format, with each file named by its extraction date (**YYYY-MM-DD.json**) and includes fields like source, URL, date extracted, title, author, and content. The dataset was last updated in August 2015 and is designed to aid research in Arabic computing or linguistics. We will be translating this to English using the googletrans Python package.
4. News Category Dataset: <https://www.kaggle.com/datasets/rmisra/news-category-dataset>
   1. The News Category Dataset on Kaggle consists of around 200,000 news headlines from the year 2012 to 2018, categorized into 41 categories. Each record includes attributes like the category, headline, authors, link, and a short description. The dataset, sourced from HuffPost, is stored in JSON format and aims to facilitate news category classification and other natural language processing tasks.
5. Apart from these datasets, we are also using beautifulsoup4 and selenium packages on python to extract article data from ndtv.com, indiatoday.in, and hindu.com—three news media websites that cover Indian news.

## 2. Retrieval of the raw data

## 1. We utilized the Kaggle api to write a script to get the news category dataset using an API call.

## 2. The SaudiNewsnet dataset is hosted on github. We used the requests package on python to access the folder containing all the data and pulled it into a given location.

## 3. As for the All the news 1.0 and 2.0 datasets, we directly downloaded the sqlite and csv files from the website.

## 4. Finally, to get data from ndtv.com, indiatoday.in and hindu.com, we plan to use beautifulsoup4 and selenium packages to scrape the data and save them as csv files.

## 3. From raw data to tidy tabular data in pandas

## By using the pandas package, we converted the various datasets into dataframes and then created corresponding csv files for each.

## For the sqlite database(All the news 1.0) we used pd.read\_sql\_query() method to convert it to a dataframe

## For the json files in Saudinewsnet and news category dataset, we used pd.read\_json() method and then concatenated all the dataframes into one prior to saving them as csv files.

## 4. Data enrichment

## Primarily, we enrich our data by obtaining certain metrics from the current data.

## For each article in our dataset, we intend to run a llm based text summarization method to obtain a short 2 sentence summary stored in a new column called “article\_summary”.

## For each news headline and article\_summary, we intend to use LLMs to obtain the corresponding sentiment and emotion exhibited.

## The sentiment can be “positive”,”Negative”, and “Neutral”

## The emotion can belong to one of 6 classes namely anger, sadness, joy, surprise, love, and disgust

## The article\_summary results are stored in new columns called “article\_summary\_emotion” and “article\_summary\_sentiment”

## The headline results are stored in new columns called “headline\_emotion” and “headline\_sentiment”

## The Saudinewsnet database is in the Arabic language, we plan to convert it to English using the googletrans package.

## For each News outlet, we will also add a total count of articles

## 5. Data cleaning

## Null values are handled by dropping the corresponding rows.

## We only select certain columns such as “date”,”headline”,”article\_text”,”author”,”news\_outlet”,”country\_of\_origin”

## We utilize our enrichment pipeline to obtain other metrics

## Column headings are renamed to follow the same naming convention (such as renaming source to news\_outlet)

## Duplicates are dropped. They are identified by looking at the headline column.

## An ID column is added to make it easier to store and retrieve data down the line.

## All the different datasets are merged into one csv file using pandas.

## 9. Computation of meaningful summary statistics

## Visualization pertaining to count of articles published over time

## A choropleth map to visualize the density of articles in our dataset across the world.

## Average(mean) number of articles per region

## A pie chart with percentage of positive, negative and neutral news items

## A donut chart indicating total number of articles for each emotion.

## Drill down capabilities

## 11. Around 4-6 (or more) visualizations related to the dataset

* Ngram(Unigram, Bigram, Trigram) visualizations using bar chart
* Semantic Networks to understand the topics covered in the dataset
* Wordmaps
* ML Model related graphs (Training accuracy vs Testing Accuracy, ROC curve)
* Subplots for each country
* Subplots for each News Outlet
* Temporal analysis of news articles in the dataset