Project of Database and Analytical Programming

Shahrukh
Master of Science in Data Analytics
National College of Ireland
Dublin, Ireland
x23135379@student.ncirl.ie

Muddassir Ahmed
Master of Science in Data Aanlytics
National College of Ireland
Dublin, Ireland
x23138688@student.ncirl.ie

***Abstract*—In the current days news article analysis after they got published is a useful technique to draw meaningful conclusions about the elements that are present in the news The results that we get from analysis provide us useful insights if they are represented graphically. The purpose of this project is to perform several database operations on the news datasets then do data transformation on it and after it providing insights to the news data . For this purpose we first collected the related dataset after it moved this dataset into Mongo DB , and then perform ETL on the collected dataset in the next step we moved both the datasets into postgresql and then finally we have performed analysis on both the datasets.**

I. Introduction

News gives us newly received information; we have to use several parameters if we want to detect the quality of news that is being spread. In this project, our motivation was to find the related datasets that could relate to each other from any perspective. It was quite a challenging task because most of the datasets that we gathered could not relate to each other in any perspective. The related datasets that we explored were mostly in structured form. For this project, the datasets we have chosen are NEWS datasets. The first data set defined the news data is collected by the web aggregator. News in the dataset falls under categories like news of business, health, and science and technology category. The second dataset is in XML format, and it is gathered via API and uploaded on the website. It was also of the domain news and this dataset defined the news of a typical case that is published in different newspapers a period. To fulfill the requirements of the project we first stored both the datasets in Mongo DB, then after performing a transformation on the datasets we stored it in PostgreSQL by accessing this transformed data we performed the word frequency analysis, time stamp analysis, and author engagement analysis on both of the datasets that provided insights into data. We have found insights on both datasets by finding the repetitive words that occur in the articles in both of the datasets when the articles in the dataset got published and who are the dominant authors in both the datasets. After it a future work suggestion is given that how machine learning could be used on both these datasets

II. Related Work

While working on our project we have gone through several articles that were considering the news articles domain area. [1], [3], [4]are visualizing and analyzing the data obtained from a social media platform named as Twitter to provide useful insights into the news that is trending at that time, the topics that are trending at that time are also covered in these papers. According to the data insights provided by [4] twitter is considered as a good platform for discussion and sharing of the latest news. [1] defined a step wise approach that involves text analysis by using natural language processing , and then after it structured analysis is also performed . In the end, visualization techniques are used to get insights about the data. [3] sentiment analysis is performed in this paper by using three ways classification of data extracted from twitter. [5] in this paper future trends that could exist in news are analyzed by using K mean machine learning algorithm. In the paper [2] KNN algorithum is applied for the classification of new cases by considering distance .

Identify applicable funding agency here. If none, delete this.

III. Methodology

A. Dataset Description

1) DataSet1: The first dataset is related to news that is covering different categories like business, health and science and technology the references to news pages from where that data is fetched is also mentioned in the URL feature of the dataset. This data is being fetched from the UCI repository and it is defining various clusters of the news articles

2) DataSet2: Data set 2 is NYSK and it's the defining collection of news in which a single case of sexual assault was discussed in the various articles published at that time and the person against which the sexual assault case was registered was a former IMF director.

B. Detailed description of data processing algorithm

In Figure 1 workflow of our project is explained with the help of a diagram that is made in the draw.io tool. The procedure from finding the data to getting the two datasets that could relate to each other. One of our data was in CSV format the other data was in XML format. We uploaded both the data into our Mongo DB the purpose of storing the data in Mongo DB is firstly it was the requirement of our project, moreover, Mongo DB provides flexibility in the data modeling and provides high performance to index any field in the document. Mongo DB is used to store data by using the least amount of memory consumption which is why it is considered efficient for storing data. While fetching the

data from the Mongo DB the step of preprocessing is done in this steps, we decided to choose only those columns that are considered compulsory for our visualization and the rest of the columns that don't seem to provide useful insights to our data are eliminated and that process is used for our data cleansing. In this step we checked the missing values in our datasets, but no missing value is detected to that's why we moved on to our next step

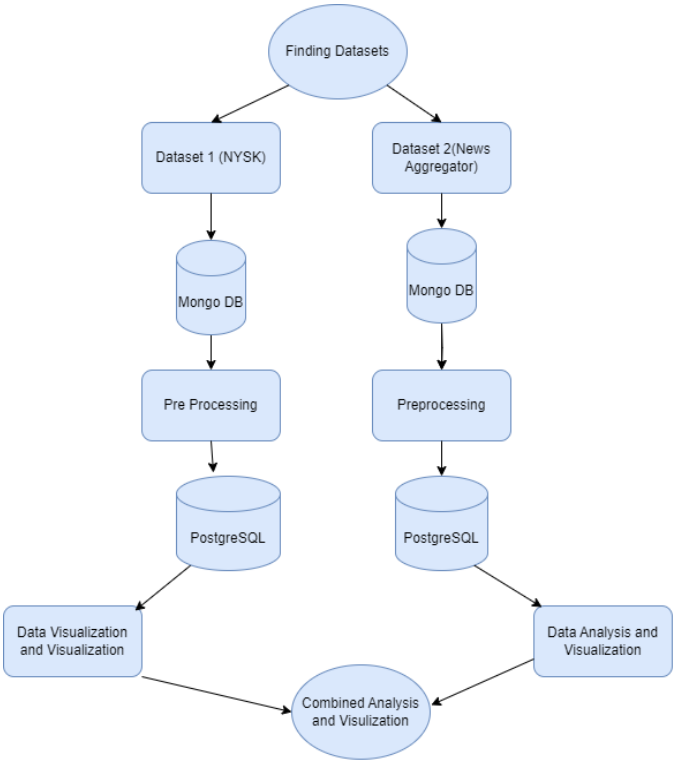


Fig. 1. Work flow

The purpose of using PostgreSQL is firstly it was the project requirement, moreover, it gives better reliability, data integrity, and better performance. In PostgreSQL, we can perform various checks for instance checking not null value. Furthermore, we can perform various SQL queries on our dataset and get meaningful information by using this.PostgreSQL is an open source database and it's the most important benefit of using it its code is freely available throughout. In fig 2 the view of the dataset that was in the XML format and contains data of the incident that got published in the different news articles is showning our first step . The data is kept in the tool Mongo DB for the further analysis

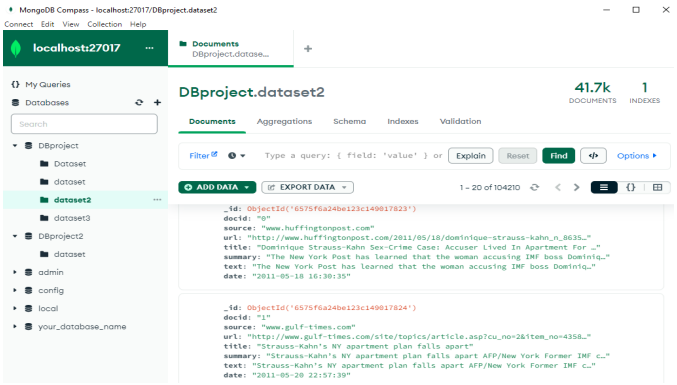


Fig. 2. DataSet2 in Mongo DB

After successfully storing dataset into mongoDB, we applied next phase which involve extracting data from mongoDB and transferring data into PostgreSQL databse. Nysk dataset name is defined for this dataset in PostgreSQL. It contains attributes such as docid, url, title, source, text and date. It is also showing all relevant data related to attributes.

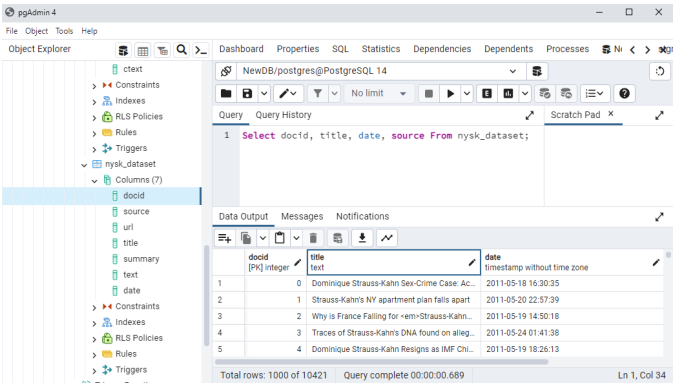


Fig. 3. DataSet1 in postgreSQL

As we have two dataset we visualized two views in the different tool . The screenshots attached above describes one dataset view in mongo db and the other dataset view in the PostgreSQL

IV. Results and Evaluation

The process of visualization is done with the help of several useful graphs. For our dataset 1 we have performed visulizations and the explanation of each visualziation is mentioned with the figure . To start up the data visualization we are checking first five rows of dataset. It includes attributes such as author, date, headlines, readmore, text and ctext. Author shows name of all authors those collected data and date show when it published, headlines show main heading of each news. Fig 4 is defining it.

```
All attributes: ['author', 'date', 'headlines', 'read_more', 'textt', 'ctext']

First 5 rows of the dataset:
             author                      date  \
0       Chhavi Tyagi  03 Aug 2017,Thursday
1         Daisy Mowke  03 Aug 2017,Thursday
2      Arshiya Chopra  03 Aug 2017,Thursday
3       Sumedha Sehra  03 Aug 2017,Thursday
4  Aarushi Maheshwari  03 Aug 2017,Thursday

                                           headlines  \
0  Daman & Diu revokes mandatory Rakshabandhan in...
1  Malaika slams user who trolled her for 'divorc...
2  'Virgin' now corrected to 'Unmarried' in IGIMS...
3  Aaj aapne pakad liya: LeT man Dujana before be...
4  Hotel staff to get training to spot signs of s...

                                           read_more  \
0  http://www.hindustantimes.com/india-news/raksh...
1  http://www.hindustantimes.com/bollywood/malaik...
2  http://www.hindustantimes.com/patna/bihar-igim...
3  http://indiatoday.intoday.in/story/abu-dujana-...
4  http://indiatoday.intoday.in/story/sex-traffic...

                                               textt  \
0  The Administration of Union Territory Daman an...
1  Malaika Arora slammed an Instagram user who tr...
2  The Indira Gandhi Institute of Medical Science...
3  Lashkar-e-Taiba's Kashmir commander Abu Dujana...
4  Hotels in Maharashtra will train their staff t...

                                               ctext
0  The Daman and Diu administration on Wednesday ...
1  From her special numbers to TV?appearances, Bo...
2  The Indira Gandhi Institute of Medical Science...
3  Lashkar-e-Taiba's Kashmir commander Abu Dujana...
4  Hotels in Mumbai and other Indian cities are t...
```

Fig. 4. Dataset 1 view

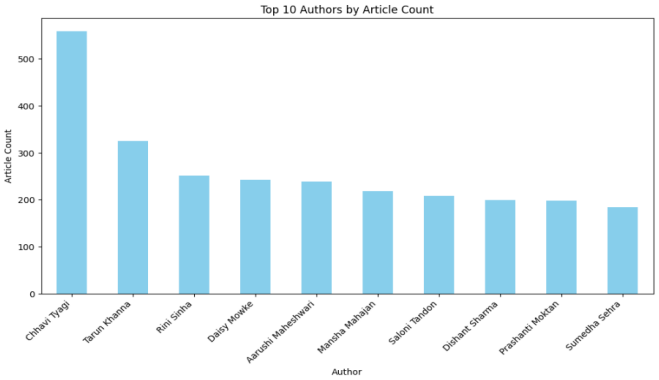


Fig. 6. Top Authours

Distribution of articles per author is with respect of number of authors shows how many authors were contributed in research. In this visualization twelve authors are the most frequent authors

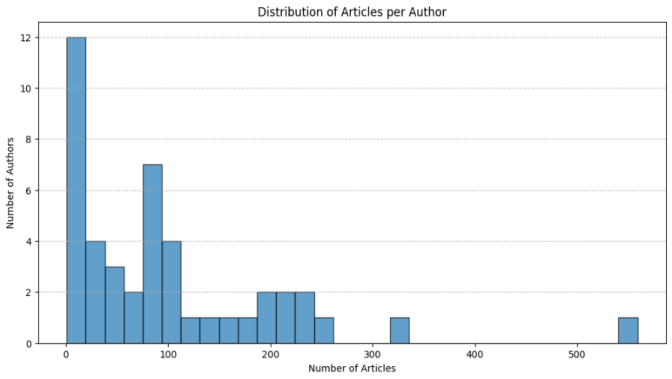


Fig. 5. Articles per authours

After this visulization the next step is to find the top authours that are publishing their articles .Pandas is used for this visualization and this dataset is composed with English news articles. Bar chart is used to show top ten authors which contributed for articles. Top ten authors by article count clarifies the purpose of visualization. X axis shows authors and y axis shows article counts of each author. This visualization is contributed in dataset for authors have contribution based on the number of articles. Chhavi Tyagi has most count of articles and sumedha sehra is at ten position of top count. Fig 6 is showing the top ten authours that are publishing articles

Aim of our project was to define the time series anlaysis that how much articles are getting published with repect to time for this purpose we have performed time series analysis with repect to the number of articles our time . The results of this are represented in the fig 7

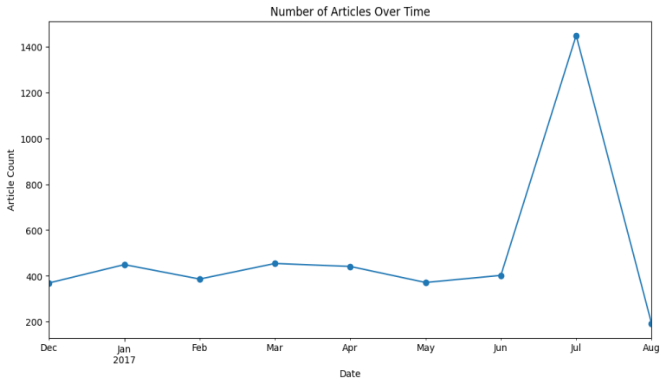


Fig. 7. Articles over time

The result of this analysis depicts that in july 2017 the maximum number of articles got published in the newspapers. If we want to apply any machine learning algorithm on the dataset headline length plays a critical role the process how we could apply machine learning to it is discussed in conclusion and future work portion . So in the visualization process we were concerned with the headline length so we have focused on it . The results can be visualized in fig 8.Histogram used to display distribution of headlines length with respect to frequency. Headline 60 shows higher than eight hundred frequencies, it is gradually increased while the headline length increased.

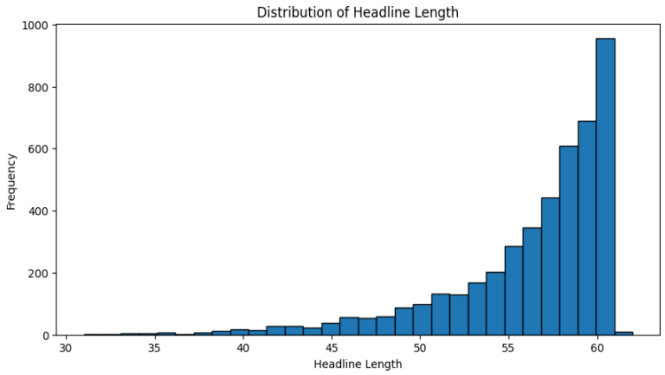


Fig. 8. Headline length

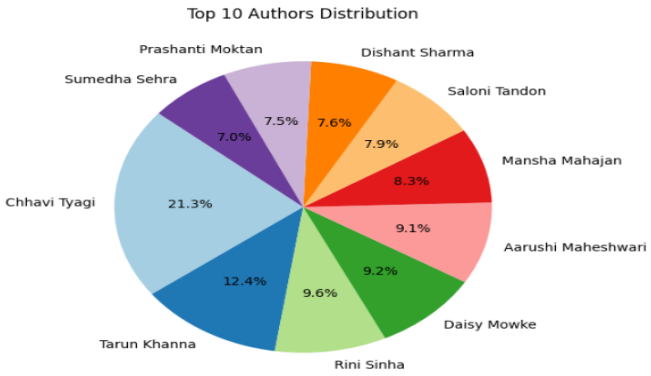


Fig. 10. Top 10 Authours distribution

We then make a word count distribution with respect to authours represented in fig 9 .Word count distribution by author is showing with respect to its news articles. Chhavi tyagi is published maximum articles as compare to all other authors. Rini Sinha is on second number to published news articles. These news articles are divided into science technology category, business category and health category.

The next step is to perform an analysis of the second dataset that was semistructured and kept in the structured form in the postgreSQL for this purpose process is involved for checking first five rows of dataset. It includes attributes such as url, title, docid, source, summary, date and text. Docid contains unique numbers, source contains where data is fetched, url contains complete website of source address, title contain subject of the story, text contains relevant data of the story and data shows when it is published . Fig 11 is elaborating this step

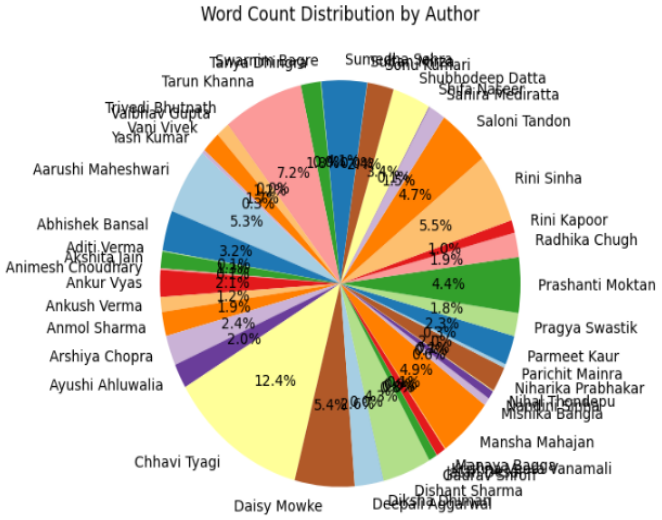
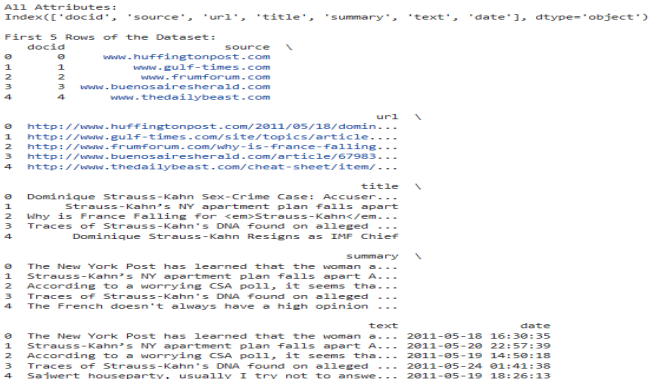


Fig. 9. Words count distrubution with news articles



```
All Attributes:
Index(['docid', 'source', 'url', 'title', 'summary', 'text', 'date'], dtype='object')

First 5 Rows of the Dataset:
   docid                source  \
0      0     www.huffingtonpost.com
1      1         www.gulf-times.com
2      2           www.frumforum.com
3      3  www.buenosairesherald.com
4      4       www.thedailybeast.com

                                                 url  \
0  http://www.huffingtonpost.com/2011/05/18/domin...
1  http://www.gulf-times.com/site/topics/article....
2  http://www.frumforum.com/why-is-france-falling...
3  http://www.buenosairesherald.com/article/67983...
4  http://www.thedailybeast.com/cheat-sheet/item/...

                                               title  \
0  Dominique Strauss-Kahn Sex-Crime Case: Accuser...
1       Strauss-Kahn's NY apartment plan falls apart
2  Why is France Falling for <em>Strauss-Kahn</em...
3  Traces of Strauss-Kahn's DNA found on alleged ...
4       Dominique Strauss-Kahn Resigns as IMF Chief

                                             summary  \
0  The New York Post has learned that the woman a...
1  Strauss-Kahn's NY apartment plan falls apart A...
2  According to a worrying CSA poll, it seems tha...
3  Traces of Strauss-Kahn's DNA found on alleged ...
4  The French doesn't always have a high opinion ...

                                                text                 date
0  The New York Post has learned that the woman a... 2011-05-18 16:30:35
1  Strauss-Kahn's NY apartment plan falls apart A... 2011-05-20 22:57:39
2  According to a worrying CSA poll, it seems tha... 2011-05-19 14:50:18
3  Traces of Strauss-Kahn's DNA found on alleged ... 2011-05-24 01:41:38
4  Salwert houseparty. usually I try not to answe... 2011-05-19 18:26:13
```

Fig. 11. Dataset 2

This graph represented in fig 10 shows rating of top distribution authors for news. Chhavi Tyagi is found most popular author for this distribution, Tarun khanan comes in second number, Rini Sinha is at third number, Daisy Mowke is at four stage and it goes on relevant to news articles.

Document Frequency by Source is indetified.The object of this visualization is showing data across different sources. X-axis: It shows document frequency by source.Y-axis: It shows number of documents representing that where news was taken and number of count documents shows how many news was published from source. Fig 12 is graphically represnting this information

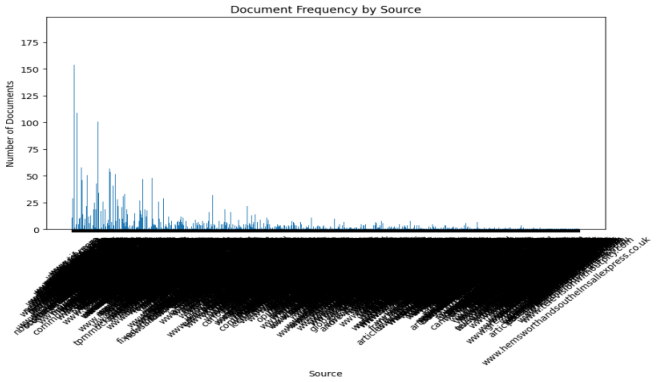


Fig. 12. Documents by source

For the time stamp analysis of this dataset two visulization are used that is depicted in fig 13 .The results of this anlaysis depicts date wise published articles. In the month of mid may, news articles were published more as compare to other date of the month.

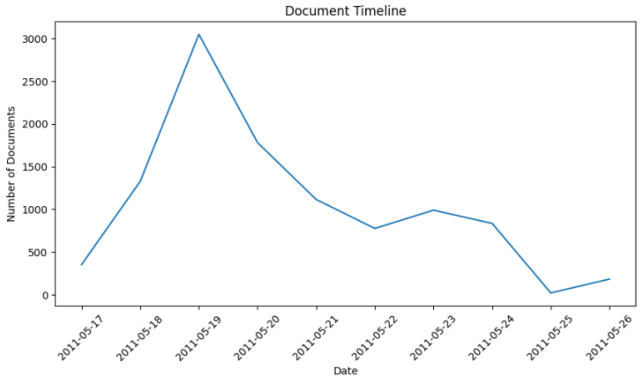


Fig. 13. Documents timeline

WordCloud for headline are used to display data of all news agencies. WordCloud use to represent text data for overview of the frequent words occurring in news headlines. It is most popular used for this purpose. In this dataset word delhi and india used most frequent in all related news. Meet, metro, top words are least count and it shows the relationship with other similar words of dataset. We have done the word frequency analysis of both the datasets as both the datasets cotain news articles so this word frequency will help us to relate these datasets the word count analysis of the dataset 1 is shown in the fig 14

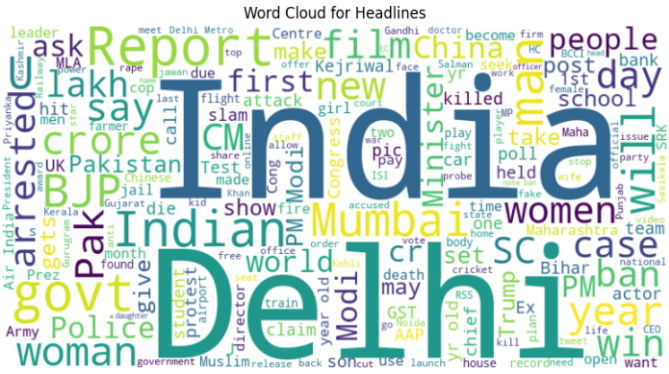


Fig. 14. Word Fequency dataset 1

Similarly, Word count analysis of our second dataset is done . The result of the secound dataset analysis is represented in fig 15. that is giving information of most frequent words.Strauss word is most commonly used in this dataset, kahn word is also showing maximum in difference articles the resason for it is that the particular person was facing the allegation

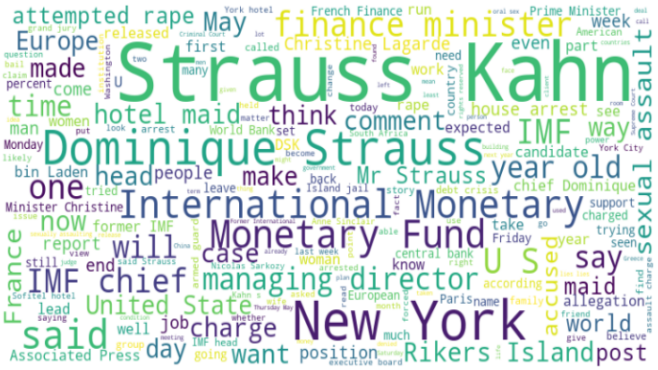


Fig. 15. Word Count Dataset 2

V. Conclusion and Future Work

In this project, we have analyzed and visualized two news-related datasets .The project is done with involving databases such as Mongo DB and postgreSQL. Time stamp, words frequency, title length, author contributions analysis is performed on the datasets and all the results are highlighted with the help of visulizations. However, more work could be performed on the provided datasets that in mentioned in the paragraph below of the future work. In the future more detailed analysis could be performed on the given datasets that could give more insights to the data. For Instance, more key words could be extracted and use to define the categories. Moreover, machine learning algorithms could be used with the dataset to define the category of news. For example, if we train a machine learning model that take title and the author's name and predict what type of news it will be or in which category this news will fall, either it will be the news related to sports, related to health care, or related to business or it is related to science and technology. We could also the give our training dataset for

the validation of our model, like , we could train our model on to our first dataset later on we use the second data set that either our first model is predicting its category in the right way

REFERENCES

[1] Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and George Giannakopoulos. An exploratory analysis of news trends on twitter. In *Proceedings of the 2016 IJCAI Workshop on Natural Language Processing meets Journalism*, pages 80–85, 2016.

[2] Padraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers-a tutorial. *ACM computing surveys (CSUR)*, 54(6):1–25, 2021.

[3] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 538–541, 2011.

[4] Rong Lu and Qing Yang. Trend analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, 2(3):327, 2012.

[5] Syafruddin Syarif et al. Trending topic prediction by optimizing k-nearest neighbor algorithm. In *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, pages 1–4. IEEE, 2017.