**Title – News Paper Analysis**

**IDA Assignment 5**

**Name – Eeshan Bablani (229309252)**

**Yuvam Kothari (229309251)**

**Class – DSE-3C**

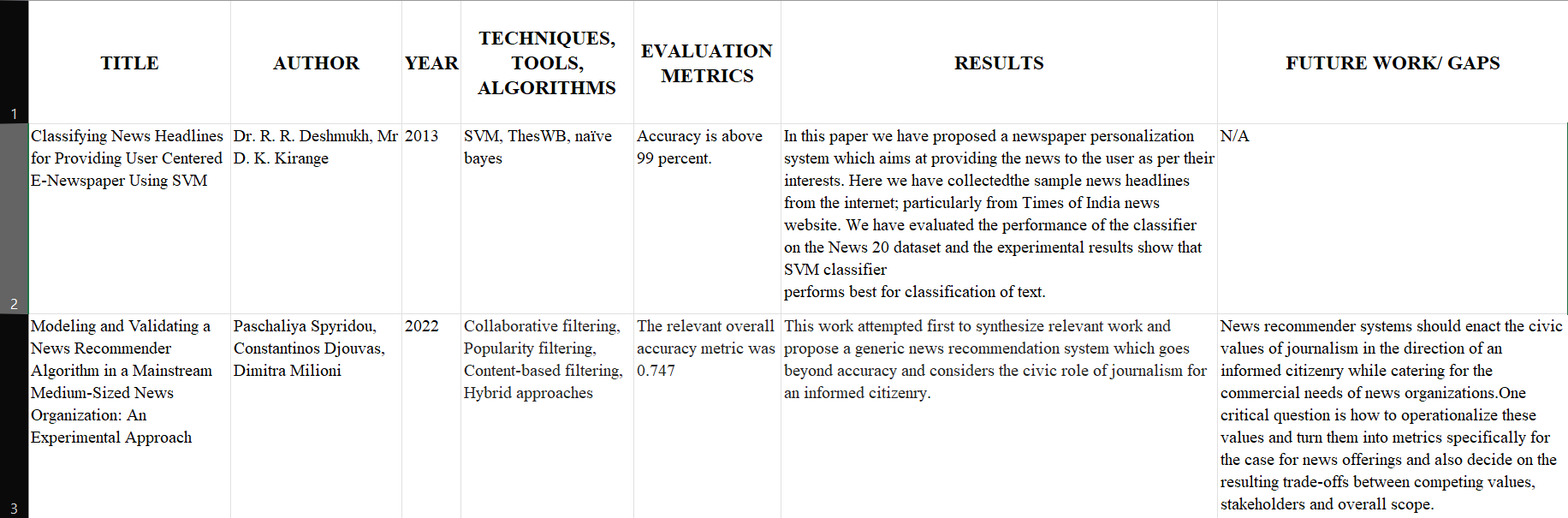
* **Abstract :**

We have been doing our work in the field of news and related commodities. In the Internet era where Social media is framing the fake news online newspaper are also not untouched. Various newspapers have been found guilty for providing fake/wrong news and advertisements in their offline/online versions, Hence we have done some assessment regarding the work that has been done using various ML/DL models and formed the Report Analysis over our project.

* + - **Literature Review :**

A literature review is a critical examination and analysis of existing work. It serves as an overview of current state of knowledge and helps us identify gaps, debates, and key themes. By evaluating previous research work, it will help us to identify areas for future work and contribute in the development of their work. We have read several articles/publications from Google Scholar and present you the literature review in a form of table, which has features such as title, authors, year of publication, results, evaluation metrics, tools and technologies used, and future work/gaps.

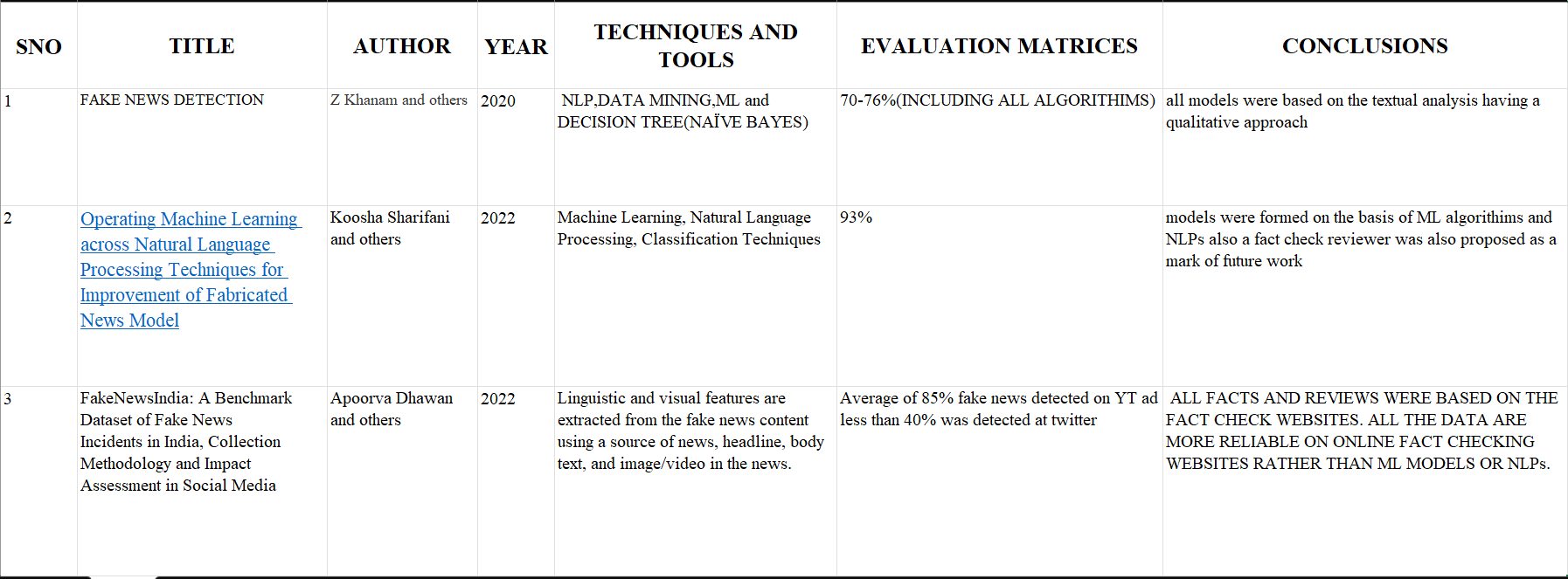
A close-up of text

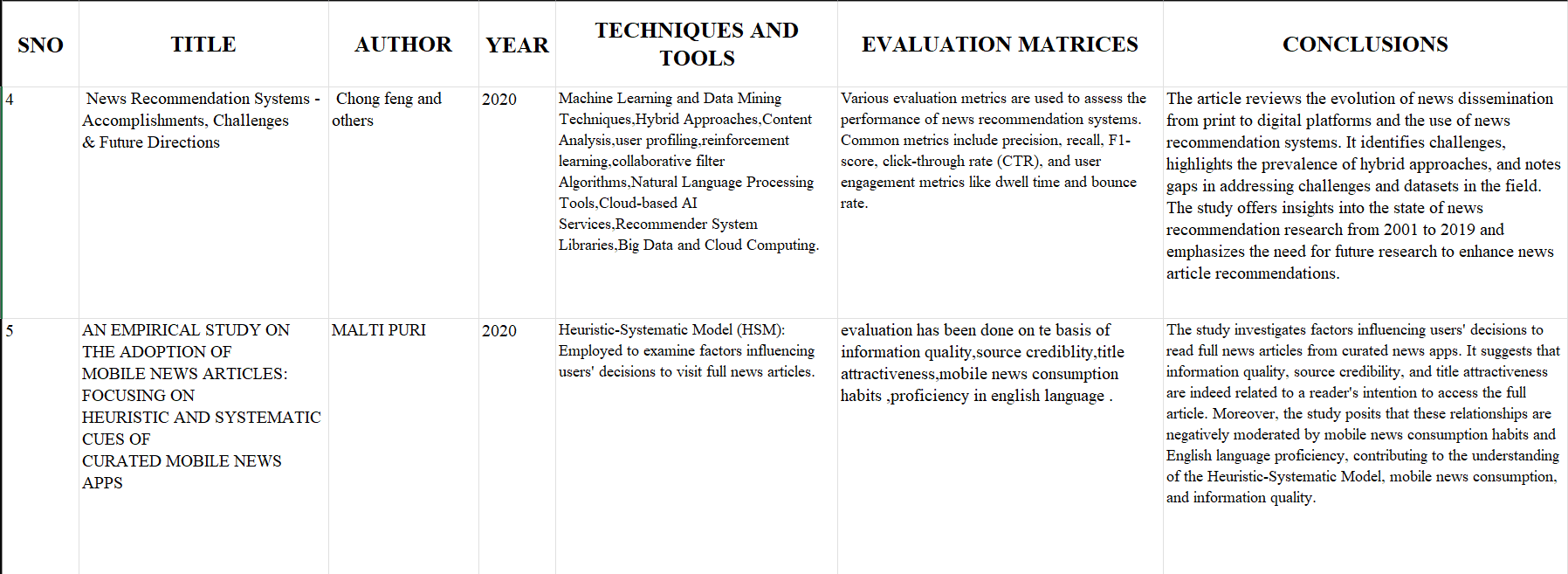
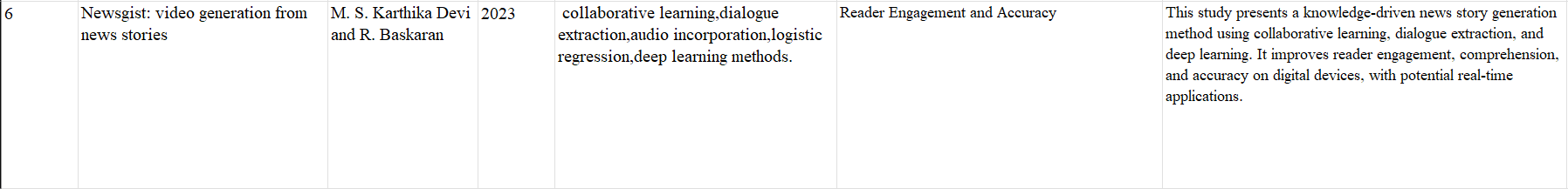
Description automatically generated

A close up of a screen

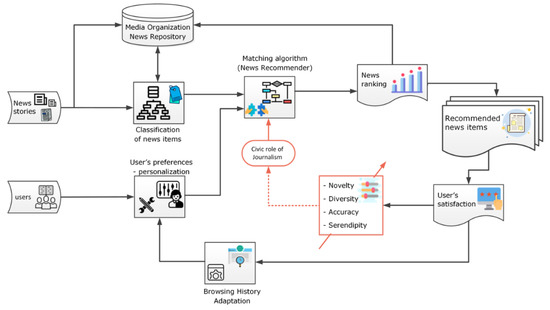
Description automatically generatedA screenshot of a computer

Description automatically generated





* **Architecture Diagram:**



News Recommender System Architecture Diagram

Reference: <https://www.mdpi.com/1999-5903/14/10/284>

A diagram of a document

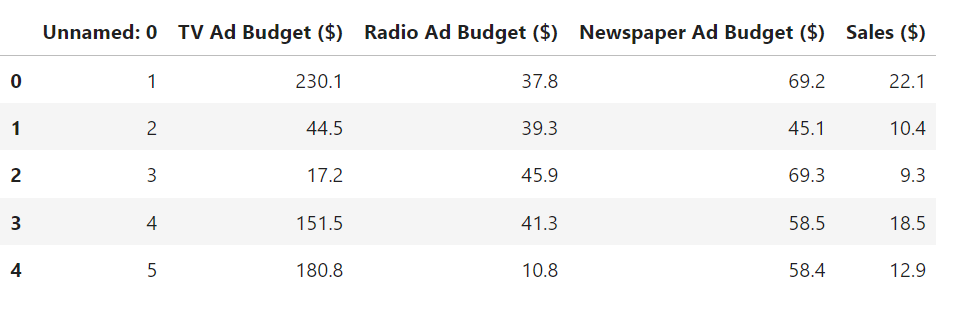
Description automatically generated

Text Categorization Using SVM

Reference: <https://link.springer.com/article/10.1007/s42452-020-2266-6>

* **Result Analysis :**

1. Advertisement’s Comparison between TV, Radio, Newspaper : In this data set , we have tried to compare the amount of the advertisement that has been displayed over TV, Newspaper and were promoted during the radio shows. We have tried to depict the insights where the regression models has been plotted using the Dependent variable as Sales and independent variable as mode of advertisement. By this we have tried to gain insights that how much people have been affected if the advertisement has found fake.



{Dataset Link: <https://www.kaggle.com/code/yasserh/advertisement-budget-prediction-top-ml-models>. }



The correlation matrix above shows that there’s a strong correlation between TV and sales (0.78), a moderate correlation between radio and sales (0.58), and a weak correlation between newspaper and sales (0.23).

A graph of blue dots

Description automatically generated

The above regression plot says that, there is a strong correlation between TV and sales as the slope of the curve is highly positive, moderate correlation between Radio and sales as the slope is positive, and weak correlation between Newspaper and sales as the slope is almost equal to zero.

1. News Classifier – The next step we had define in this project is to determine the number of spaces and letters used in the news and tried to roughly match with the original and trusted source of the news. We have Machine learning models and a primary Data set. The model has a net accuracy of 88.07%.

Dataset Link - <https://www.kaggle.com/code/hamzamanssor/news-classification-ml-models/notebook>

A screenshot of a phone

Description automatically generated

**EDA & DATA VISULAISATION**

A graph of different colored bars

Description automatically generated

The above plot says that most news in the newspaper is on technology, and least news is printed on business.

**PLOTTING DIFFERENT CATEGORIES OF NEWS**

A graph of a graph of a graph

Description automatically generated with medium confidence

The x-axis represents the values of 'Text\_length', which is a measure of the length of the text.

The y-axis represents the frequency (or count) of occurrences for each bin. Each bin represents a range of ‘20 Text\_length' values.

**PLOT TEXT LENGTH**

A graph of a number of people

Description automatically generated with medium confidence

**PLOT WORD COUNT**

A graph of a graph

Description automatically generated with medium confidence

**PLOT PUNCTUATION COUNT**

A graph of a number of objects

Description automatically generated with medium confidence

Red means Politics, Grey means Technology, Green means Entertainment, Black is Business.

In Fig. 1, x-axis: Represents Text Length of story

y-axis: represents frequency(or count) of occurrences of each bin i.e range of ‘50’.

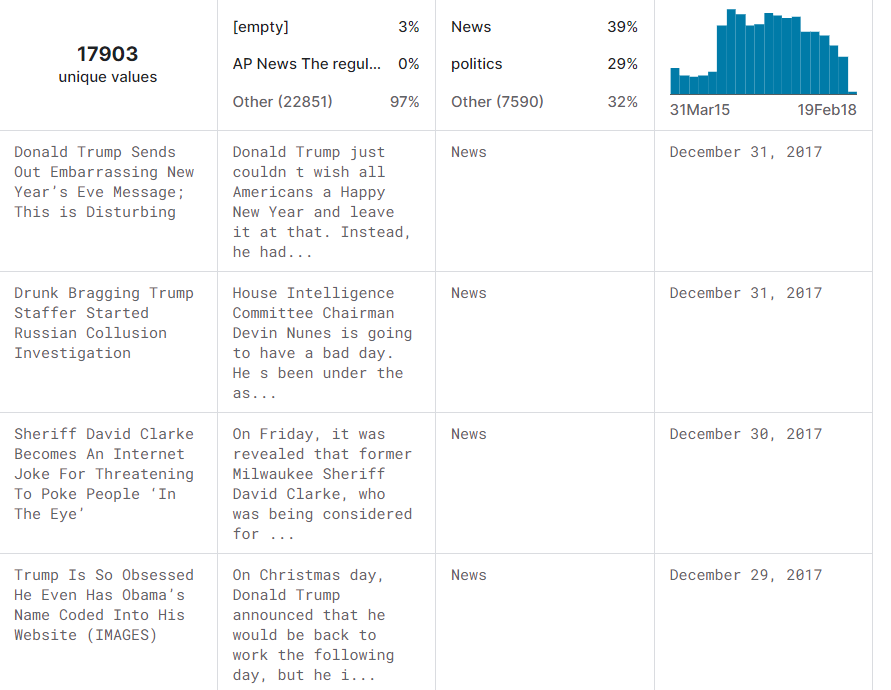
In Fig. 2, x-axis: Represents Word Count of story

y-axis: represents frequency(or count) of occurrences of each bin i.e range of ‘50’.

In Fig. 3, x-axis: Represents Punctuation Count of story

y-axis: represents frequency(or count) of occurrences of each bin i.e range of ‘50’.

1. Fake news classifier – In this method we had tried to classified news as fake or not Using the data set as well as the deep learning model for classifying the news as fake or real. The accuracy of the above model was came out to be more than 99%.



A screenshot of a graph

Description automatically generated

Data Set Link – <https://www.kaggle.com/code/karlentsatinyan/fake-news-classifier-with-accuracy-99/input?select=Fake.csv>

* **Conclusion:**

From our work we conclude the ways in which we can Identify the ways by which the public can be manipulated by fake news/Advertisements. In conclusion to our work we would like to depict that above methods can be applied and a proper fact checking website can be made were these models may be used and a genuine review can be there where the news Publishers are providing the proper news or not.

* **Future Work:**

Furthermore, we can develop a News Recommender system aimed at improving the overall news consumption experience for users. This can be achieved by reducing exposure to fake news and false advertisements. For this, we would need to analyse the amount of visualizations, fake news, repetitive news, effective news, and advertisements across different newspaper mediums and compare them. A Personalized recommender system can also be developed based on the user's language and interests. The benefits of this system include saving users time and enhancing their knowledge in preferred topics. Users can also explore different fields.

* **References:**
* <https://www.kaggle.com/code/yasserh/advertisement-budget-prediction-top-ml-models>
* <https://www.kaggle.com/code/hamzamanssor/news-classification-ml-models>
* <https://www.kaggle.com/code/kumudchauhan/fake-news-analysis-and-classification>
* <https://www.kaggle.com/code/karlentsatinyan/fake-news-classifier-with-accuracy-99/notebook>
* <https://www.kaggle.com/code/yasserh/advertisement-budget-prediction-top-ml-models>
* <https://www.mdpi.com/1999-5903/14/10/284>
* <https://www.researchgate.net/profile/Ratnadeep-Deshmukh-2/publication/262793684_Classifying_News_Headlines_for_Providing_User_Centered_E-Newspaper_Using_SVM/links/0a85e538dc9485b5d3000000/Classifying-News-Headlines-for-Providing-User-Centered-E-Newspaper-Using-SVM.pdf>
* <https://link.springer.com/article/10.1007/s42452-020-2266-6>

Links of the research papers – 1)<https://www.sciencedirect.com/science/article/pii/S0306457318306794>

2) <https://link.springer.com/chapter/10.1007/978-3-031-10539-5_14>

3) <https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/meta>

4) <https://www.tandfonline.com/doi/abs/10.1080/17512786.2014.882056>

5) <https://ieeexplore.ieee.org/abstract/document/8963698/>

6) <https://ora.ox.ac.uk/objects/uuid:f028366e-4a28-4bb1-bcfb-5ecc8f7eb45a>

7) <https://www.tandfonline.com/doi/abs/10.1080/00051144.2023.2241774>