SIES Graduate School of Technology
RISE WITH EDUCATION

SIES Graduate School of Technology
Information Technology Department AY 2022-23
Realtime News Analysis using Natural Language Processing

Group No.: 5
Group Members : Aakansha Ramesh, Swaranjali Jadhav, Gauri Thube
Project Guide : Prof. Seema Redekar

Abstract

Fake news has become a huge issue; it is both a societal & technical issue. It is difficult for companies to identify as the term covers different meanings like satire, false tales, factual errors, misleading headlines, & propaganda. The proposed application will analyze the news & classify it into Real/Fake & Clickbait/Non-Clickbait. In addition to analyzing the news into different categories, it summarizes & includes a dynamic feature that fetches live news.

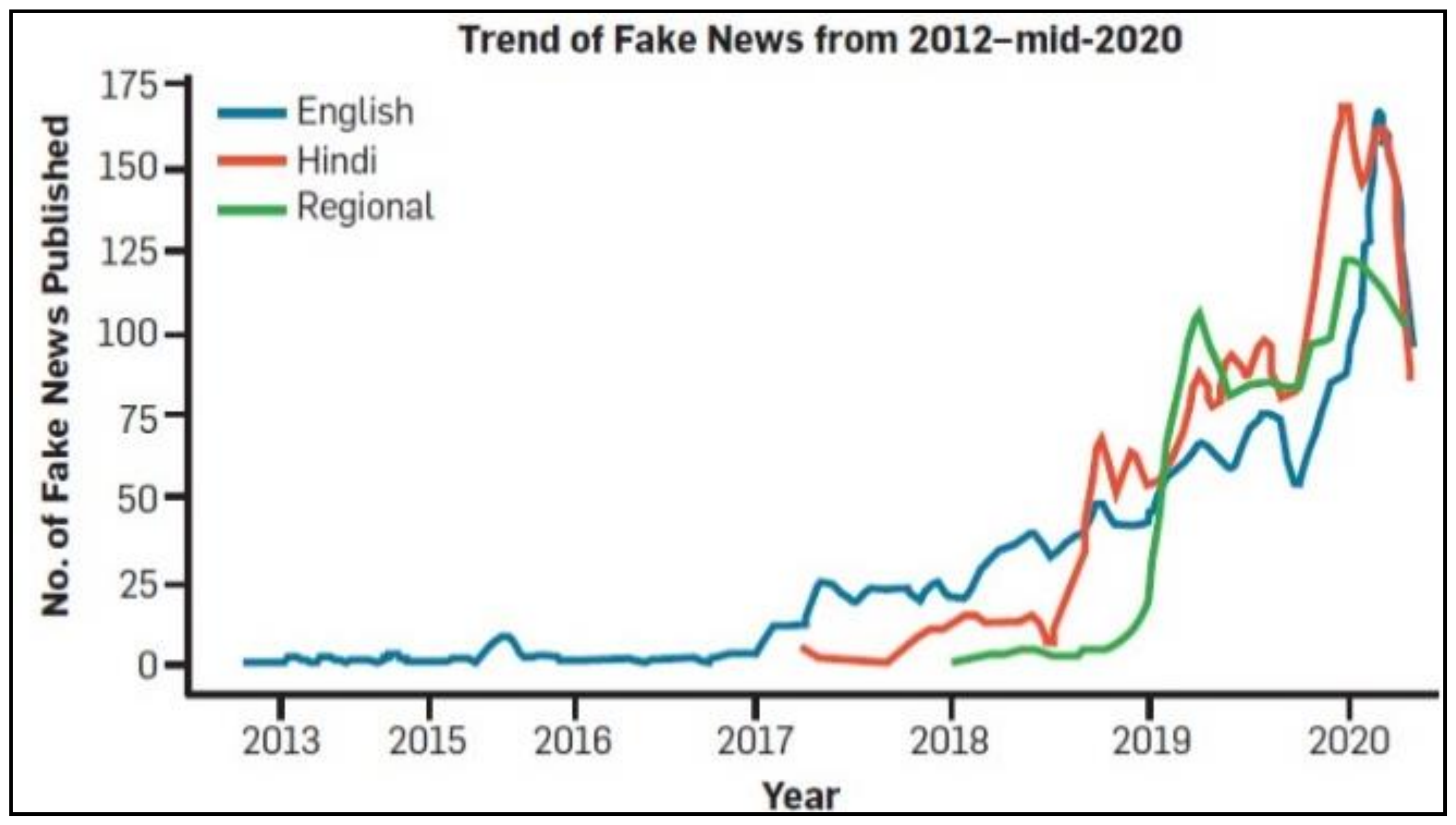
Objectives

➢The question arises about the authenticity of different news articles, especially on social networking sites where there is no check on fake news.

➢Many organizations tend to create news articles with clickbait headlines to increase their revenue.

Introduction

Transmission & information sharing are possible from anywhere across the globe. Anyone can publish content on the web today irrespective of credibility. Fake news has a large market, as it entices people & uses sensational news headlines to attract attention. The main objective is to detect fake news, which is a classic text classification problem & can be solved using Machine Learning & text processing techniques. A model needs to be built that can differentiate between "Real" news & "Fake" news & also differentiate between "Clickbait" & "Non-clickbait" news. The proposed system helps to identify the authenticity of the news.

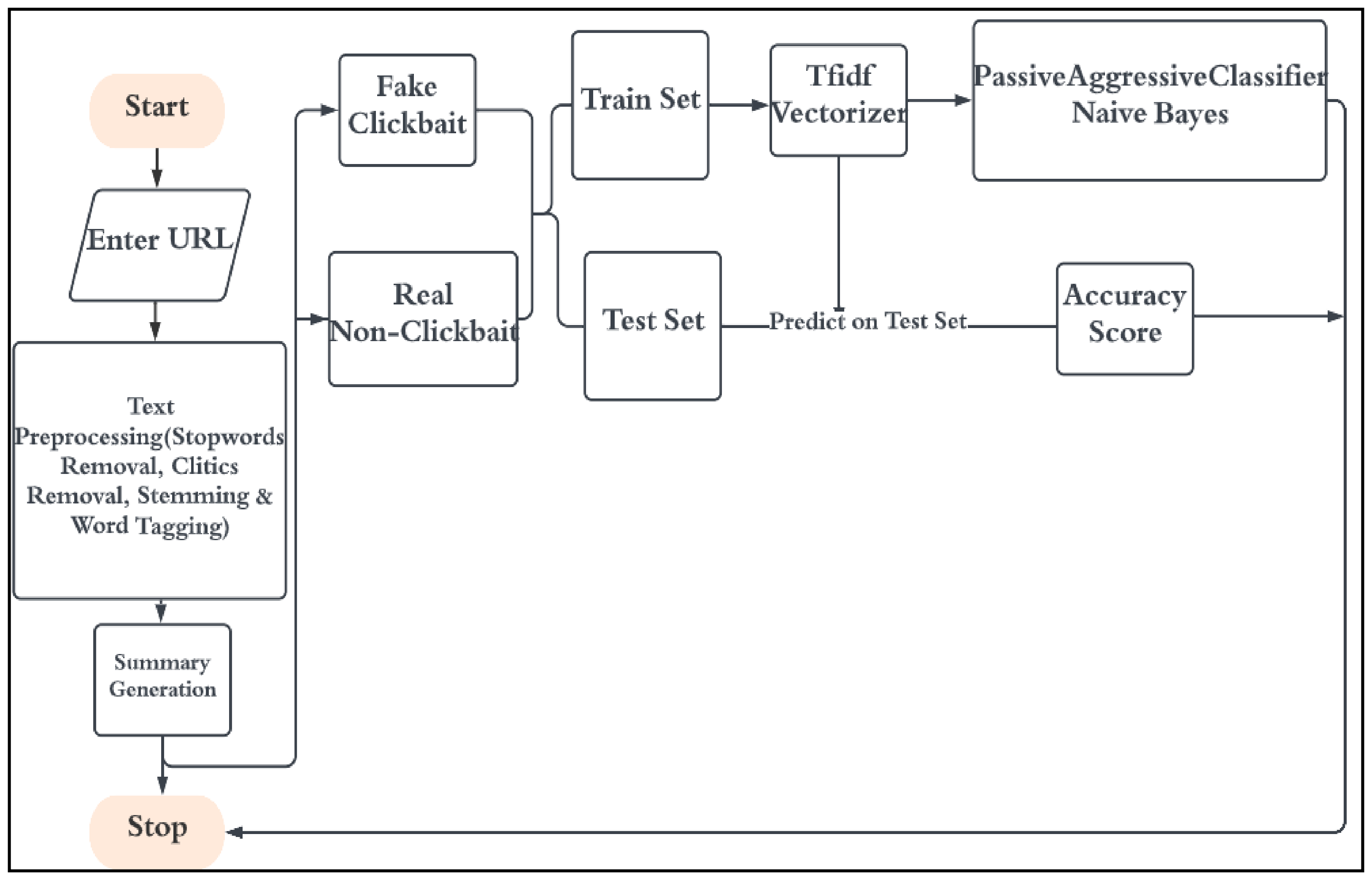


Applications

- Enables a wide range of users to stay informed.
- Reduces the risk of people being misled by incorrect information.
- Eliminates the need to review entire news article.
- Aid differently abled audience.

Methodology

- The process starts when the user enters the URL of any article & various NLP techniques like stop word removal, clitics removal, stemming & word tagging.
- Web scraping is used to fetch the data & after applying feature extraction a summarized view of the article in form of title & headline is generated.
- Machine Learning algorithms such as Naive Bayes & Passive Aggressive Classifier are used to classify the news article into Real/Fake & Clickbait/Non-Clickbait.



Functionality of System

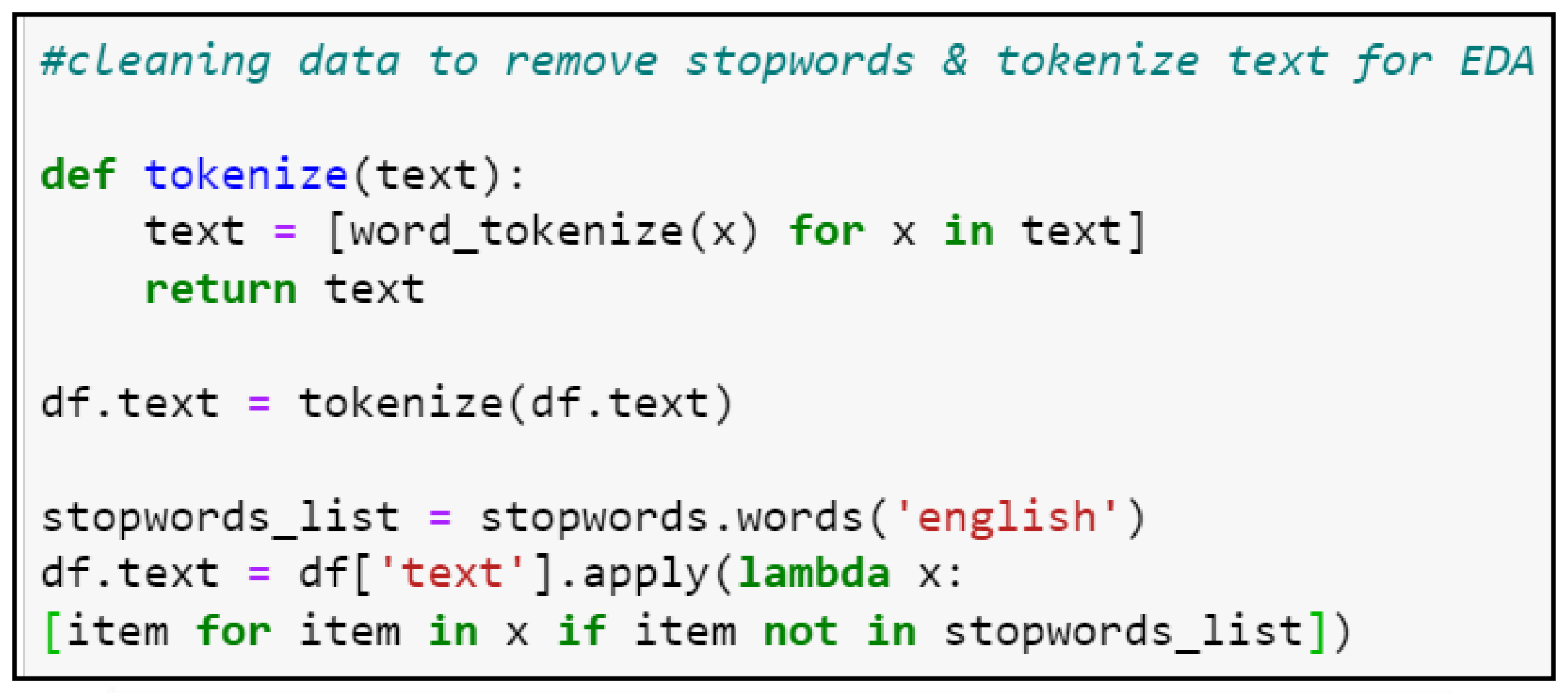
Web Scraping & Data Collection:

- The Clickbait dataset was collected from BuzzFeed, Upworthy, Thatscoop, Viralstories etc.
- The Non-Clickbait dataset was collected from NY times & the Washington post etc.
- Each of the attributes represents the web page information such as the title of shared news, the news web address, the source file of the shared news.

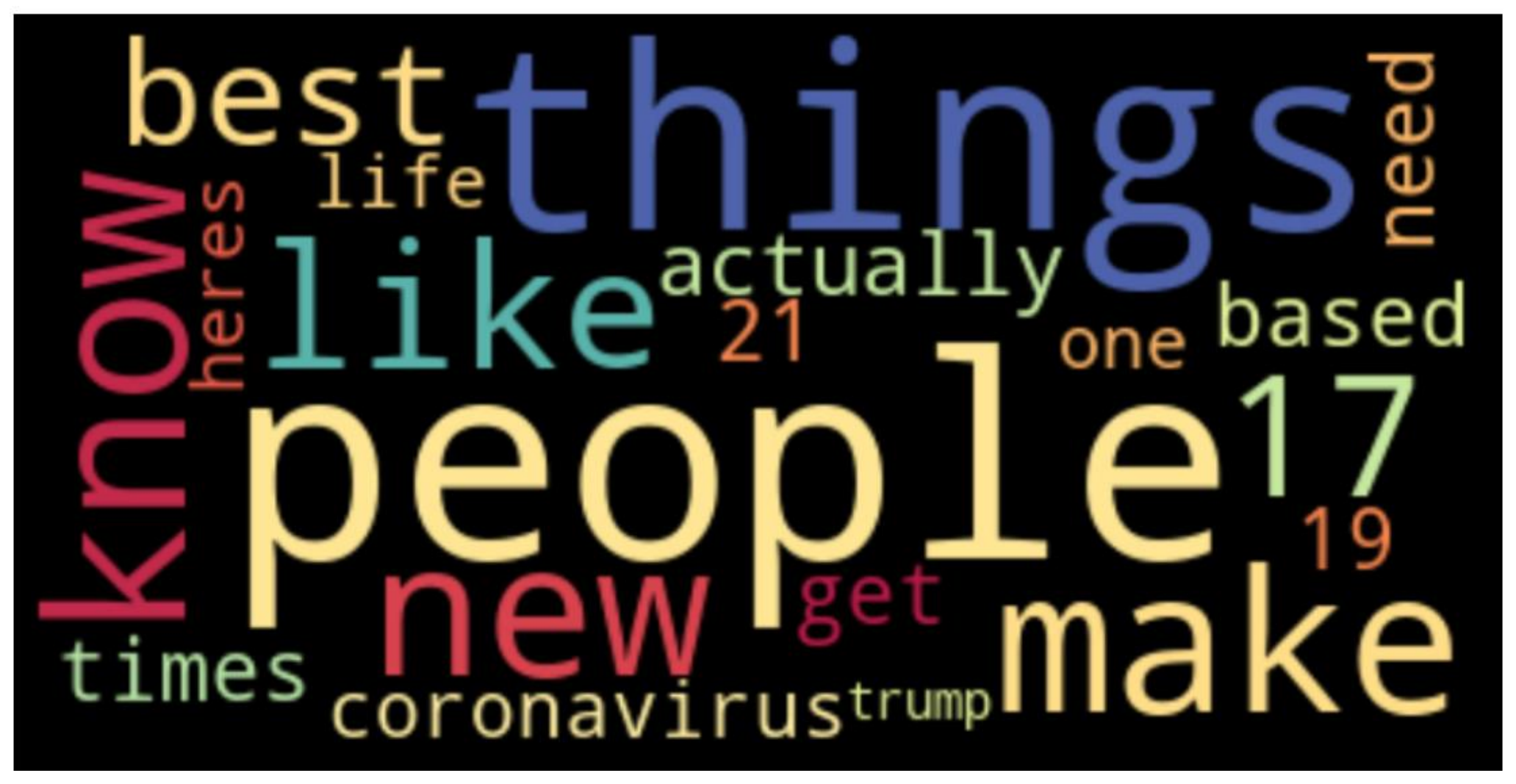
Sr no	Datasets		
	Sources	*Dimension*	*Description*
1.	Clickbait and Non clickbait sources	52,000 headlines	The headlines dated from 2007-2020. Various clickbait sources used in dataset are Buzzfeed, Upworthy, ThatScoop, ViralStories. And non-clickbait sources are from NY times, the Washington post, the Guardian Boomberg and Reuters.
2.	Kaggle Dataset	30,000 data	The data was collected from 2007-2016
3.	Twitter APIs and online publications	22,000 headlines	Headlines were web scraped from Internet.

Pre-Processing & Feature Extraction:

- NLTK was used to tokenize the body & headline of any article.
- The body & headline of the article is tokenized with Punkt statement tokenizer.
- A word cloud is created for the headline & body present in the dataset for gaining insights from the data.

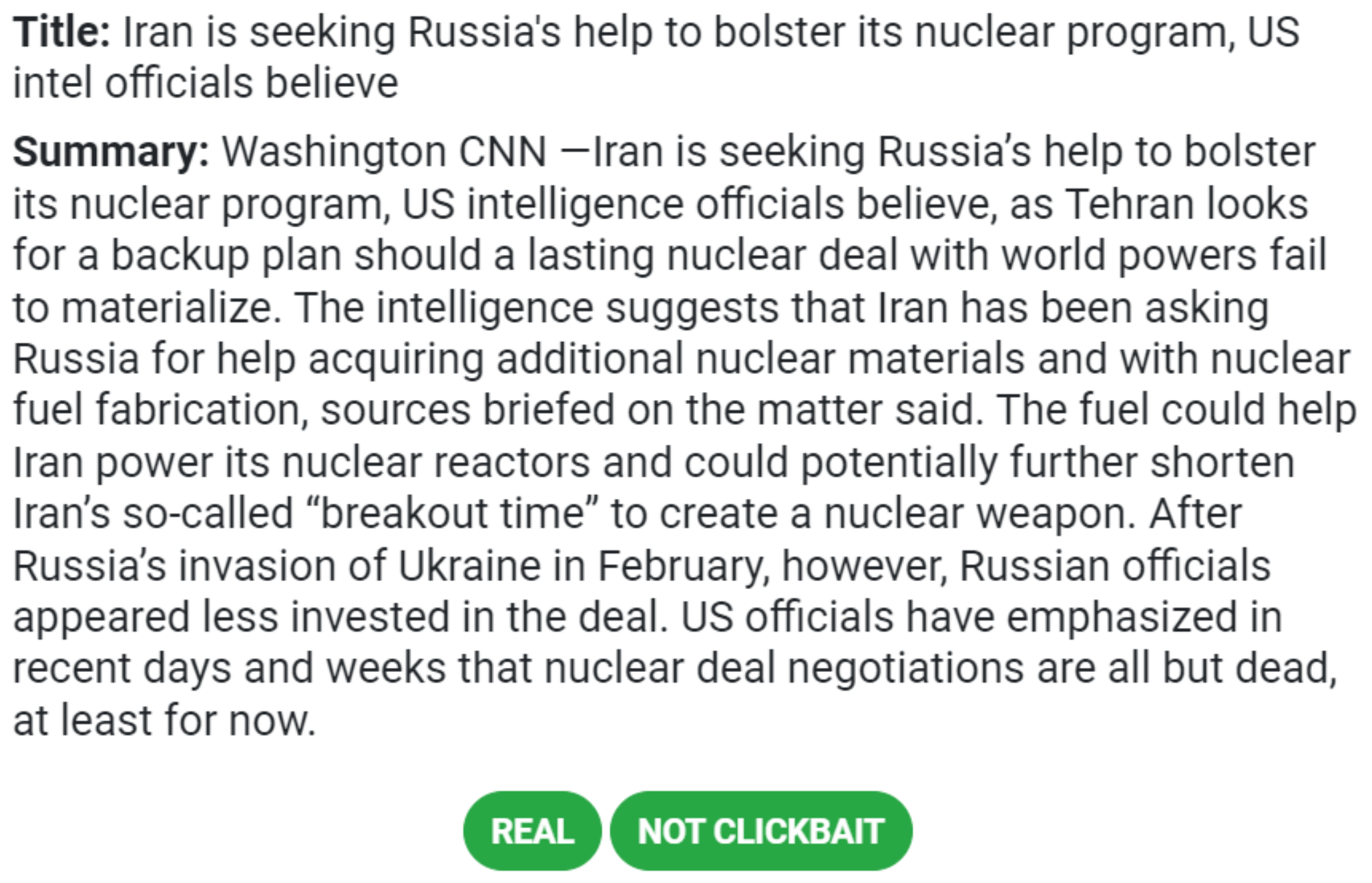


```
#cleaning data to remove stopwords & tokenize text for EDA

def tokenize(text):
    text = [word_tokenize(x) for x in text]
    return text

df.text = tokenize(df.text)

stopwords_list = stopwords.words('english')
df.text = df['text'].apply(lambda x:
[item for item in x if item not in stopwords_list])
```



Extractive Summarization:

- As, in the case of abstractive summarization, inferences are made beyond the scope of the information provided to it. Thus, extractive summarization is used

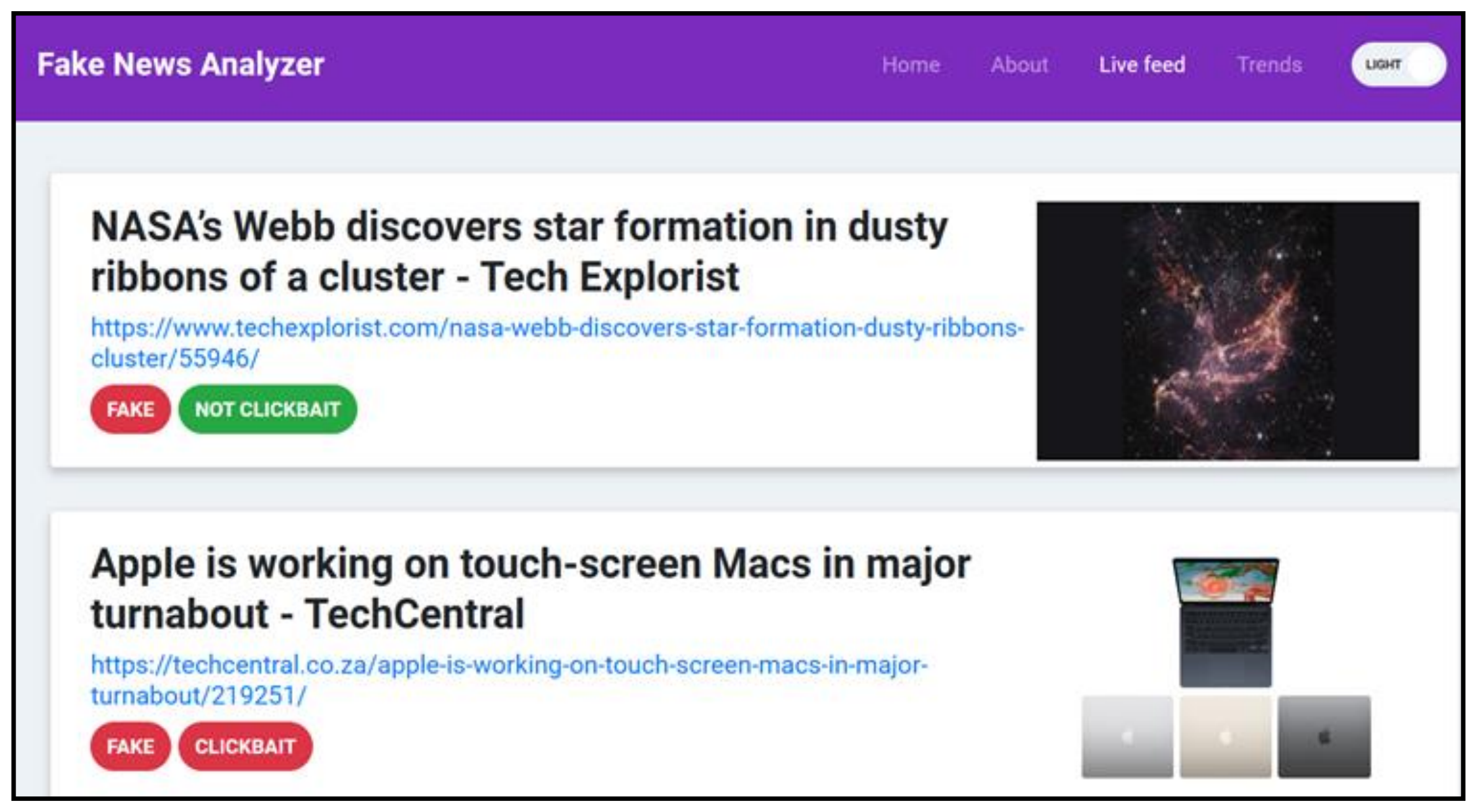


Title: Iran is seeking Russia's help to bolster its nuclear program, US intel officials believe

Summary: Washington CNN —Iran is seeking Russia's help to bolster its nuclear program, US intelligence officials believe, as Tehran looks for a backup plan should a lasting nuclear deal with world powers fail to materialize. The intelligence suggests that Iran has been asking Russia for help acquiring additional nuclear materials and with nuclear fuel fabrication, sources briefed on the matter said. The fuel could help Iran power its nuclear reactors and could potentially further shorten Iran's so-called "breakout time" to create a nuclear weapon. After Russia's invasion of Ukraine in February, however, Russian officials appeared less invested in the deal. US officials have emphasized in recent days and weeks that nuclear deal negotiations are all but dead, at least for now.

Identifying & labelling latest news:

- The application provides real-time news using the Live-Feed section & provides labels for latest news.
- News is updated on regular basis after being fetched from News API & is also tagged with the aforementioned labels.

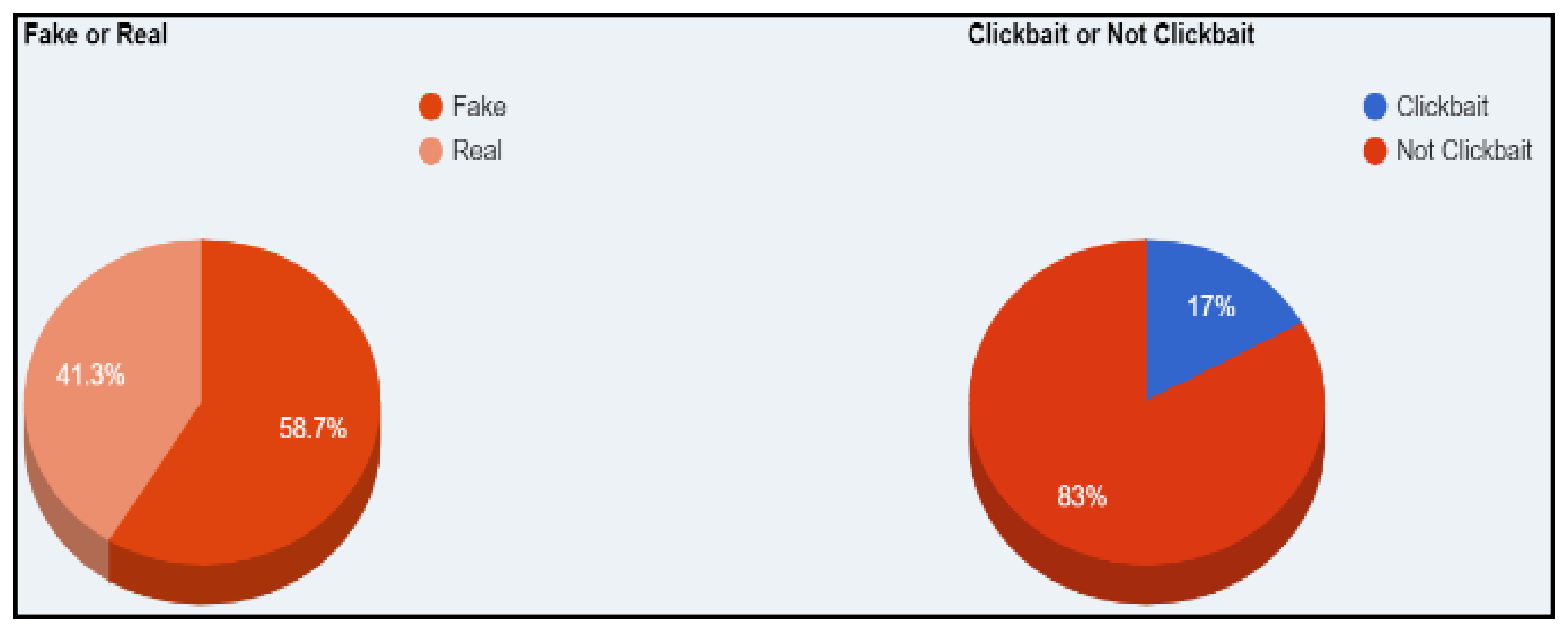


NASA's Webb discovers star formation in dusty ribbons of a cluster - Tech Explorist
https://www.techexplorist.com/nasa-webb-discovers-star-formation-dusty-ribbons-cluster/55946/
FAKE NOT CLICKBAIT

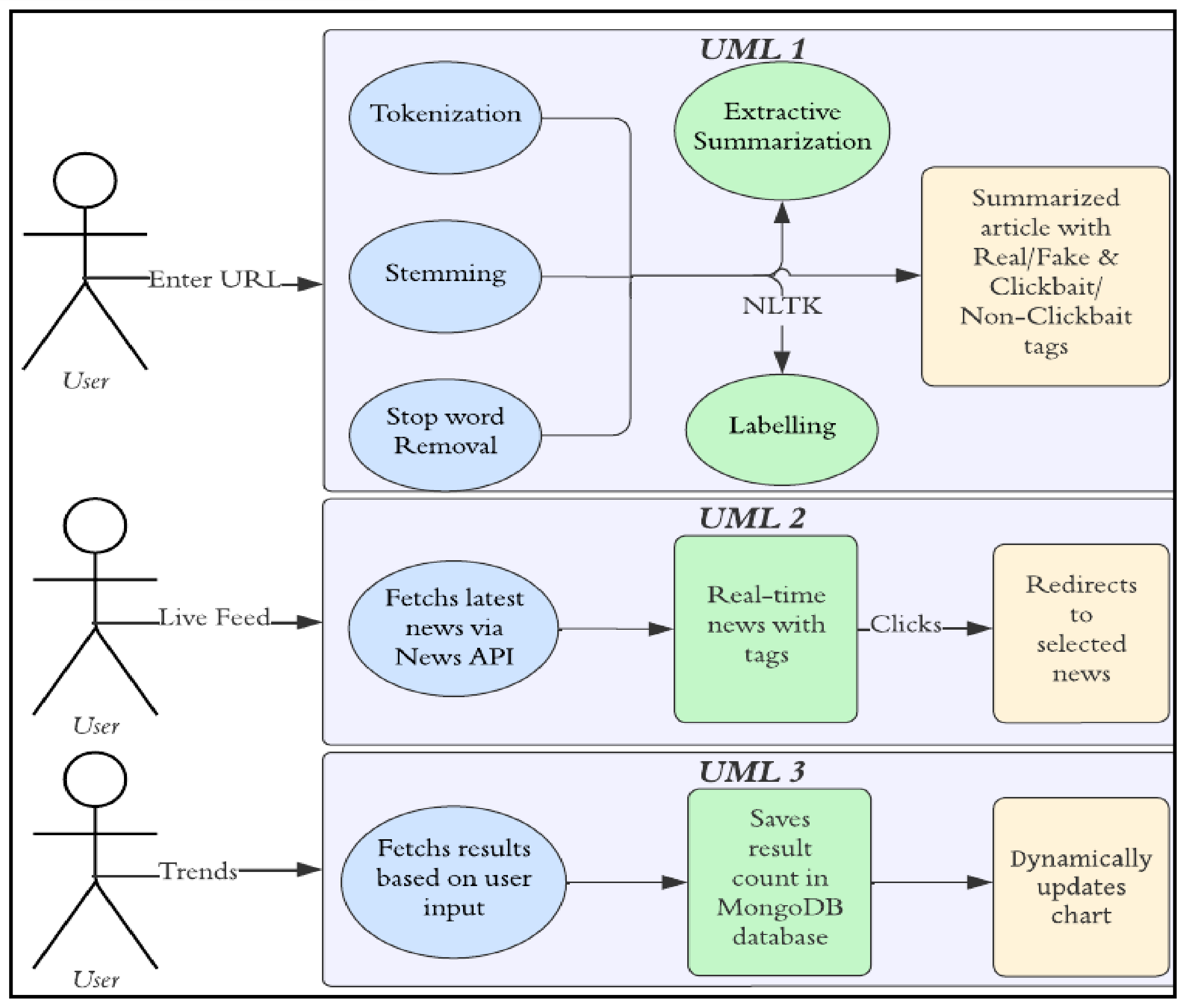
Apple is working on touch-screen Macs in major turnabout - TechCentral
https://techcentral.co.za/apple-is-working-on-touch-screen-macs-in-major-turnabout/219251/
FAKE CLICKBAIT

Statistics of Application:

- Trends Page contains the statistics, wherein Google charts drawing feature in JavaScript is used to display the results in the form of a pie diagram.
- The displayed results have the classified outputs that will be saved in the MonogoDB database.
- The database stores the count of the URL submissions based on the results & uses this count to compare the different categories.



UML Diagram:


UML 1
User
Enter URL
Tokenization
Stemming
Stop word Removal
NLTK
Extractive Summarization
Labelling
Summarized article with Real/Fake & Clickbait/ Non-Clickbait tags
UML 2
User
Live Feed
Fetchs latest news via News API
Real-time news with tags
Clicks
Redirects to selected news
UML 3
User
Trends
Fetchs results based on user input
Saves result count in MongoDB database
Dynamically updates chart

Results & Discussions:

- We have performed hyperparameter tuning for Clickbait & Non-Clickbait news, algorithms such as Naive Bayes, SVM & Logistic Regression were used.
- An accuracy ranging between 90% & 93% was observed.
- For classification of Real or Fake news Multinomial Naive Bayes & Passive Aggressive classifiers were used, accounting with an accuracy of 88% & 91.9% respectively.

Conclusion

Fake news sharing is one of the popular research problems in recent technology based on lack of security and trust in terms of the truth of shared news in social media. NLP has revolutionized different sectors & has a tangible impact on the detection of fake news. However, most existing systems still have issues and fails to meet user expectations. Thus, our application resolves the conundrum whether supposedly credible news sites are trustworthy.

Publications

[1] Accepted in IEEE conference, 4th INTERNATIONAL CONFERENCE OF EMERGING TECHNOLOGIES 2023.

References

[1] Tandoc Jr, Edson C. "The facts of fake news: A research review." Sociology Compass 13, no. 9 (2019): e12724

[2] Naeem, Salman Bin, Rubina Bhatti, and Aqsa Khan. "An exploration of how fake news is taking over social media and putting public health at risk." Health Information & Libraries Journal 38, no. 2 (2021): 143-149.

[3] Pennycook, Gordon, and David G. Rand. "The psychology of fake news." Trends in cognitive sciences 25, no. 5 (2021): 388-402.

[4] Bryanov, Kirill, and Victoria Vziatysheva. "Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news." PLoS one 16, no. 6 (2021): e0253717.

[5] Gupta, Saloni, and Priyanka Meel. "Fake news detection using passive- aggressive classifier." In Inventive Communication and Computational Technologies, pp. 155-164. Springer, Singapore, 2021.