Grenze International Journal of Engineering and Technology, Jan Issue

NEWSIFY- Categorisation and Perception of Suspicious News using Machine Learning Algorithm- A Survey

Yash Gujar[1], Rohit Shelar[2], Pratiksha Londhe[3], Niranjan Pawal[4] and Sonali Rangdale[5]
[1-5]JSPM'S Rajarshi Shahu College of Engineering, Pune, MH-India, Department of Information Technology
Email: yashgujar14@gmail.com

***Abstract*—Reading news articles is an integral part of our daily life. Journalists and editors are tasked with determining which articles are popular so they can efficiently allocate resources to support a better reading experience. This paper aims to achieve a model which simultaneously performs all the three aspects i.e: detecting whether the given news article is fake or not, summarizing a given news article and predicting the popularity and ranking the news article. The reasons behind the popularity of news articles tend to be varied and may include contemporaneity, writing quality, and other potential factors. In this article, we consider the problem of popularity prediction as a regression, developing several classes of features (metadata, contextual or content-based, temporal and social) and building models to predict popularity. The system presented here will be used in a real time environment. The other aspect of the paper is a summary of news articles on a particular topic. Since there are also various online resources that publish fake news, it is difficult to judge and understand which articles are fake. This paper focuses on the classification of the news as fake or real. The different Natural Language Processing and Machine Learning algorithms are studied for the implementation of the model. This paper helps the users for the identification of real and fake news which are published online. This will satisfy the problems of users willing to obtain quality news content in a short period of time.**

***Index Terms*— Deep Learning, Machine learning, Fake news detection, Newsify, Sentiments, Summarization of article, Popularity Prediction.**

I. INTRODUCTION

News is an important part of every individual's daily life. Each and every day of our life is incomplete without NEWS. As a human, one can easily identify real news from fake news with outlandish source of the information piece. In this paper we are primarily focusing on three aspects related to News which include: Fake News Detection, Summarizing the News Article and Finding out the popularity of news article.

Fake News is one of the most important problems in today's world of glamour, where celebrities and athletes are often made to face with rumours about their life. Many of these people regularly face this problem and find ways to overcome it. The second part of our paper is to summarize a given piece of news article. This is one of the most important things in today's world where people are in constant hurry of finishing off things quickly. So people can go through this summary and get knowledge about the news article. The third and final part of this

Grenze ID: 01.GIJET.9.1.534_4

paper is about finding the popularity of the given news article. People nowadays tend to follow trends in everything they do. Hence news is nothing different from all this. People always show interest in trending news and try to follow that. So this can also be accomplished using our methodologies.
To put it briefly, this article offers consumers a platform where they may access a combination of all the aforementioned elements in one location. People can select any of the offered items based on their personal preferences and have delight in reading or analysing the news.
Is there a specific time of day when people look for news that goes beyond the headlines? In general, 4 out of 10 people claim to have read more about a certain news topic than just the headlines in the previous week. When they did, their in-depth reading, viewing, or listening followed a pattern similar to how people typically consume news, with the majority (34 percent) stating that they don't have a preferred time to read in-depth news. This data calls into question the idea that despite seeing headlines constantly, people save their evenings for learning more.

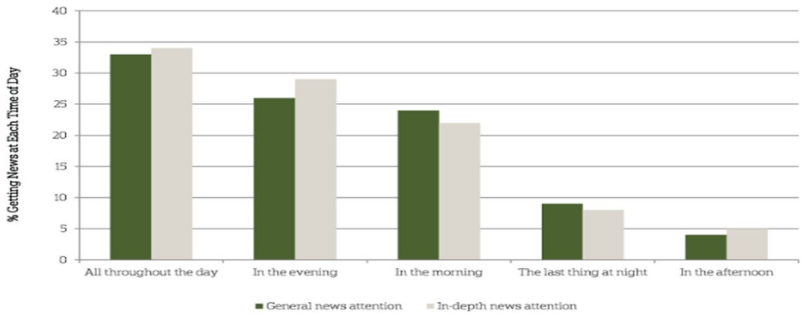

Fig-1: News Attention throughout the day

Additionally, although the time of day was not questioned, a significantly higher percentage of adults, 49 percent, stated they investigated further to discover more about the most recent breaking news story they paid attention to.

II. Review of Literature

A. Architecture

The issue of long-distance dependence arises when the input is a lengthy document since the basic single-layer encoder-decoder architecture is unable to adequately capture the relationship between the contexts while encoding the document.
The image describes the architecture of the process of summarizing a text article. It comprises of various stages. At first, the dataset provides the model with data for preprocessing . In this stage the data undergoes various operations such as splitting, tokenization, distributed control and word embedding.

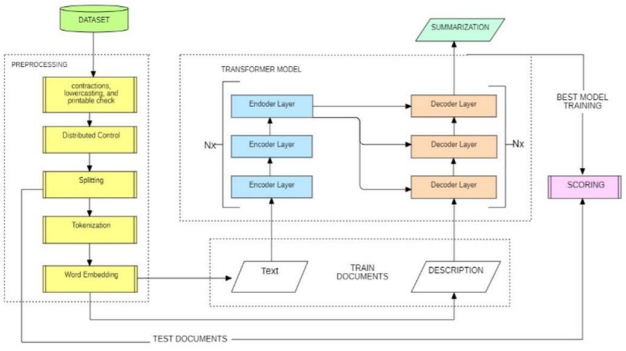

Fig-2: Architecture for summarization of text

Let us get some more knowledge about this below:
Splitting: When data is divided into two or more subgroups, this is known as data splitting. A two-part split often involves testing or evaluating the data in one part and training the model in the other. Data splitting is a crucial component in data science, especially when building models from data. This method aids in ensuring the

accuracy of data model construction and the processes that use data models, such as machine learning. There are various types of splitting like random splitting, pre-splitting, sequential splitting.
Tokenization: Tokenization is a simple process that converts unprocessed input into a meaningful data string. Even while tokenization is most known for its use in cybersecurity and the creation of NFTs, it is an essential stage in the NLP process. Tokenization is a method used in natural language processing to separate sentences and paragraphs into smaller, more easily understood parts.
Word Embedding: When words are represented as real-valued vectors for text analysis, it is referred to as word embedding in NLP. A development in NLP has enhanced computers' capacity to comprehend text-based content more effectively. It is regarded as one of the most important developments in deep learning for tackling difficult natural language processing issues. This method represents words and documents as numeric vectors, allowing words with similar meanings to have comparable vector representations. In order to deal with text data and maintain the semantic and syntactic information, the collected characteristics are put into a machine learning model. Once translated, this information is employed by NLP algorithms that are simple to process.
Distributed Control: This refers to distributing the work flow amongst different modules.
After then, control moves to the level of the transformer model, where data is encoded and decoded. During this procedure, many types of text data are trained, which aids in summarising the content. We only use GELU once on each encoder or decoder layer in the feed-forward network because of its high complexity in the field of NLP, but the performance it generates is superior to that of other activation functions like ELU and ReLU.
The proposed system uses the commented dataset as input and related information such as date, source, and author. Then it converts them into usable feature datasets used during the learning stage. This conversion is preprocessing, it performs a series of operations such as clean up, filtering, and encoding. The preprocessed dataset is Divided into two parts, one for training and the second for testing. The training engine uses the training data set, support for vector machine algorithms for building decision models.
This can be applied to the test data set. If the model is adopted (i.e. it can achieve an acceptable rate of accuracy), then it can be kept and used till the training is over. Otherwise, the parameters of the learning algorithm are adjusted and the accuracy is improved.

A. Algorithms

OCL and PUL algorithms for fake news detection: - One-Class Learning (OCL) algorithms learn from instances with a single class labelled on them, typically treating the interest class as a positive class.

i. Naive Bayes Classifier: -

The Nave Bayes algorithm is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in text categorization with a large training set. One of the most simple and effective classification algorithms now in use is the Naive Bayes Classifier. It facilitates the creation of efficient machine learning models that are capable of making precise predictions.
Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur.
Spam filtration, Sentimental analysis, and article classification are a few examples of Naive Bayes algorithms that are frequently used.
Consider the classification procedure of the naive Bayes classifier. Given that the words of the news items being detected are being iterated over, it is possible that some specific terms won't even show up in the training dataset. In these situations, the current method just ignores these phrases, classifying them as neither Real nor false. This is a serious weakness in this model. When utilising this strategy with social media platforms like Facebook and Twitter, a minor tweak is required. If certain phrases are absent from the training data gathered using the web scraping technique, we will mark it as "FAKE."

ii. K Means Clustering algorithm: -

K-Means Clustering is a type of unsupervised learning algorithm that is used to address clustering issues in data science or machine learning. In this chapter, we will learn what the K-means clustering method is, how it operates, and how to implement it in Python.It gives us the ability to divide the data into various groups and provides a practical method for automatically identifying the groups in the unlabelled dataset without the need for any training.
The method makes use of a database of articles that were found to be false news by employing a programme called the BS detector, which essentially maintains a blacklist of websites that publish false news.
The dataset includes articles and metadata from 244 different websites, which is advantageous in that it will prevent the model from picking up source bias due to the variety of sources. However, it is clear from a quick

glance at the dataset that there are still some straightforward explanations for which a model may discover details about what is contained in the "body" text in this dataset.

iii. Support vector machine algorithm: -

One of the most widely used methods for Supervised Learning, Support Vector Machine (SVM), is used to solve Classification and Regression issues. In machine learning, it is generally employed to solve classification issues.The SVM algorithm aims to construct the optimum decision boundary or line that can divide n-dimensional space into classes so that we can quickly classify fresh data points in the future. Known as a hyperplane, this optimal decision boundary.

The extreme vectors and points that help create the hyperplane are chosen via SVM. The SVM approach is based on support vectors, which are utilised to represent these extreme situations. View the diagram below to see how two different categories are categorised using a decision boundary or hyperplane.

They choose the support vector machine algorithm to train the model. This enables the decision function value provided for a news item to be used as a measure of the confidence in that item's classification: a positive decision function value designates both a true news item and the level of truth it possesses, and vice versa; a negative decision function value designates both a fake news item and the level of misleading content it possesses.

iv. Latent Semantic Analysis:

Latent Semantic Analysis is a technique for examining a collection of documents in order to find statistically significant word co-occurrences that reveal information about the themes on which those words and documents are used. It is a technique for natural language processing that examines connections between a collection of texts and the terms used therein. It scans unstructured material using the mathematical approach of singular value decomposition to look for undiscovered connections between phrases and ideas. Concept searching and automatic document classification are the two main uses of LSA. However, it has also been used in search engine optimization and publishing (text summarising). In the end, LSA reformulates text data in terms of r latent features, also known as hidden features, where r is smaller than m, the total number of terms in the data.

v. Lexical Analysis:

Lexical analysis is the process of attempting to comprehend the meaning of words, infer their context, and take note of the relationships between various words. It frequently serves as the beginning of many NLP data pipelines. There are numerous different types and variations of lexical analysis. It takes a source code file and converts the lines of code into a collection of "tokens," deleting any whitespace or comments. For example, it is used as the initial stage of a compiler. Lexical analysis may preserve multiple words as a "n-gram" in other types of analysis (or a sequence of items). After tokenization, the computer will look up terms in a dictionary and attempt to determine their meanings. For a compiler, this would require finding keywords and associating operations or variables with the keywords. In other cases, such as in a chat bot, the lookup may require accessing a database to match intent. The meaning of a word must be determined by the context in which it is used because, as was already mentioned, a word may have multiple meanings.

vi. LexRank:

LexRank is an unsupervised method for text summarization that scores phrases based on their centrality on graphs. The key concept is that sentences "recommend" other sentences that are similar to them to the reader. As a result, if a sentence sounds a lot like many others, it probably has a lot of significance.

The significance of the sentences "recommending" it also contributes to the significance of this sentence. Because of this, a statement needs to be similar to many other sentences in order to rank highly and be included in a summary. This makes logical sense and enables the algorithms to be used on any fresh text at random.

vii. Predicting the Popularity of Online News from Content Metadata:

Popularity of a news article is important to rank the news articles according to their popularity. We can prioritize the news articles on the basis of the number of views it gets in the first 24 hours after its publication. After that we apply the regression technique and extract the features like social, temporal and contextual for forecasting and then we get the page view count. Furthermore, to get more accuracy in predicting the popularity of news article, we can also add the features like number of repost, number of likes, number of comments, etc. Popularity of a news article can be predicted in two ways which are: Before Publication and After Publication.

Before publication is quite easier because we have the required data available like number of views, likes. But before publication is complex because we need to predict whether the given post will get reposted or not on the basis of meta data.This paper predicts that the user will share this article or not, using gradient boosting machine

using features that are known before publication. This GBM technique uses the gain of contribution of each feature towards the popularity prediction and then computes the mean of it. The main point to note here is that GBM predicts the popularity using only statistical data of news and without using after publication content.

TABLE I: COMPARISON OF RESULTS

Sr No	Paper Title	Year	Author	Accuracy
1	Centrality Revisited for Unsupervised Summarization	2019	Hao Zheng, Mirella Lapata	84.7%
2	Automatic text summarization of news articles	2017	Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R. B. Keskar	95%
3	Predicting the Popularity of News Articles.	2016	Y.Keneshloo, S.Wang, Eui-Hong (Sam) Han, Naren Ramakrishnan	85%
4	Predicting online news popularity based on machine learning	2022	Min-JenTsai, You-QingWu	88%
5	Predicting and Evaluating the Online News Popularity based on Random Forest	2021	Yuchao Zhang, Kun Lin	93.4%
6	Predicting the popularity of online news from content metadata	2020	Md. Taufeeq Uddin; Muhammed Jamshed Alam Patwary; Tanveer Ahsan	84%
7	Fake News Detection Using Machine Learning Algorithms	2020	Uma Sharma, Siddhart Saran, Shankar M. Patil	91.65%
8	Fake News Detection Using Machine Learning	2021	Vijaya Balpande, Kasturi Baswe, Kajol Somaiya, Achal Dhande, Prajwal Mire	91%
9	Fake News Detection Using Machine Learning Approaches	2021	Z Khanam, B N Alwasel , H Sirafi, M Rashid	96%
10	Research On The Prediction Of Popularity Of News Dissemination Public Opinion Based On Data Mining	2022	Jing Tian, Huayin Fan, And Zengwen Hou	91%
11	News Text Summarization Based on Multi-Feature And Fuzzy Logic	2020	Z Khanam , B N Alwasel, H Sirafi and M Rashid	93.43%

III. CONCLUSIONS

Many researches going on all over the world emphasizing on various points in a news or information snippet to identify and inform of any fake news. Consuming news articles is an integral part of our daily lives and various news agencies expend tremendous effort in providing high quality reading experiences for their readers. The uniqueness of the proposed system lies in the fact that currently there isn't a model which would provide users with a combination of all the things that are: identifying the fake news, summarizing the news article as well as predicting it's popularity. After the survey of machine learning algorithm, a detailed information about various machine learning algorithms that would be used in all the three aspects of our project in order to auspiciously implement this project was determined. In this paper firstly a system has been developed in order to identify whether a given news article is fake or not using various machine learning algorithms. In above subsequent approach, later the problem popularity prediction problem as regression is resolved using different metadata techniques. Knowledge From above results we could conclude that different types of existing methods that can be used for fake news identification. This paper also successfully summarizes the news article which saves the time of user and gives us a short description about news article.

REFERENCES

[1] Laxmi Rananavare, P. Venkata Subba Reddy, "Automatic News Article Summarization", 2018
[2] Hao Zheng, Mirella Lapata, Sentence "Centrality Revisited for Unsupervised Summarization",2019.
[3] Prakhar Sethi, Sameer Sonawane, Saumitra Khanwalker, R. B. Keskar," Automatic text summarization of news articles",2017.
[4] Muhammad Mohsin, Shazad Latif, Muhammad Haneef, Usman Tariq, Muhammad Attique Khan," Improved Text Summarization of News Articles Using GA-HC and PSO-HC", 2021
[5] Prof.Asha Rose Thomas, Prof.Teena George, Prof. Sreeresmi T S," Automatic Text Summarization of News Articles",2020.
[6] Nouf Ibrahim Altmami, Mohamed El Bachir Menai," Automatic summarization of scientific articles: A survey,2022.
[7] Y.Keneshloo, S.Wang, Eui-Hong (Sam) Han, Naren Ramakrishnan,"Predicting the Popularity of News Articles.", November 2016.
[8] M.J.A Patwary, T.K. Ahsan "Predicting the Popularity of Online News from Content Metadata", October 2020.
[9] Author panelMin-JenTsaiYou-QingWu," Predicting online news popularity based on machine learning", 2022.

[10] Yuchao Zhang1 and Kun Lin2," Predicting and Evaluating the Online News Popularity based on Random Forest",2021.
[11] Md. Taufeeq Uddin; Muhammed Jamshed Alam Patwary; Tanveer Ahsan,"Predicting the popularity of online news from content metadata",2020.
[12] Akshay Jain, Amey Kasbe, "Fake News Detection", 2018.
[13] Nihel Fatima Baarir**,** Abdelhamid Djeffal, "Fake News detection Using Machine Learning", 2020.
[14] N S S Ramachandra , S Sandeep, B V Kishore, "Fake News Detection Using NLP", 2021
[15] Uma Sharma, Siddhart Saran, Shankar M. Patil, "Fake News Detection Using Machine Learning Algorithms", 2020.
[16] Vijaya Balpande, Kasturi Baswe, Kajol Somaiya, Achal Dhande, Prajwal Mire, "Fake News Detection Using Machine Learning ", 2021.
[17] Z Khanam , B N Alwasel , H Sirafi, M Rashid, "Fake News Detection Using Machine Learning Approaches", 2021.
[18] Anjali Jain, Harsh Khatter, Avinash Shakya, "A Smart System for Fake News Detection Using Machine Learning", 2019.
[19] Jing Tian, Huayin Fan, And Zengwen Hou, "Research on The Prediction of Popularity of News Dissemination Public Opinion Based on Data Mining", 2022
[20] YAN DU, HUA HUO, "News Text Summarization Based on Multi-Feature and Fuzzy Logic", 2020.