Fake News Predictor: A Random Forest-Based Web Application for the Prediction of Fake News on Social Media

Aritra Basu, Cinu C. Kiliroor, Ritam Basu, and Ritabrata Nag

Abstract The fake news is one of the main concerns nowadays. There are different groups who spread fake news and use that to gain popularity or defame others. In the domain of fake news analysis, our main focus is to help the users to understand whether news is true or fake. The existing models present in the market analyze news based on the text or some time by fact checking in the Internet. These types of model are very time-consuming. Our model is designed based on the identification of the pattern of comment, reaction, and share count on any news. Based on these parameters the proposed model predicts whether the news is real or fake.

Keywords Streamlit · Random forest · News type · Fake news

1 Introduction

Social media is becoming a strong media platform and day by day more are using social media. People are using it for different purpose starting from chatting, sharing info, job searching, the list is endless [1]. One of the biggest uses of Facebook is to get news of the world and people are browsing through the Facebook news, rather than watching TV [5]. Even nowadays mainstream media are also posting news on their Facebook pages. But one problem with this thing is that we cannot verify the news shared or published by the media [1]. So, many times it happens that news shared are not only fake but also cause violence. Even media are also sharing news which is half-true or politically biased or completely fake. So this causes problem for common people [2]. So, the question genuinely rose how much we can rely on Facebook news [4]. Our fake news analysis is based on the following parameters: comment, share, and reaction count, news agencies which has published. Based on these parameters it can predict whether news is real or fake.

The main objective of the proposed model is to get an idea of how social media is becoming a new weapon for propaganda and a hub for fake news [10]. Fake news spread thousands of times faster than the real one and the fake news spreaders use

A. Basu · C. C. Kiliroor (✉) · R. Basu · R. Nag
SCOPE, Vellore Institute of Technology, Chennai, India

social media as a platform for fake news to spread. News generally shared on social media are half-true or completely false and politically motivated [7]. To understand how fake news controlling our day-to-day activities and our lifestyle, everyday many events are happening throughout the world like protest, election, new technology, etc. [8]. The way these things should be presented is now in the hand of social media and the media. Often these things are represented in a way that will benefit them.

2   Literature Review

With the growth in spread of false information through the social media platforms, it is of necessary importance for individuals and societies to be able to judge the authenticity of it. Fake news is not a new trend it has been used from very early to spread propaganda, but it is a commonly occurring phenomenon in recent days. There are many ways to identify the wrong information. By conducting a systematic literature review, authors identified main approaches currently available to identify fake news and how these approaches can be applied and modified for different situations in different situations [5]. Fake news detection topic has gained a great deal of interest from researchers around the world [7]. Many people are more involving in this topic and got researcher's attention [9]. Discussion on news events has become a routine of people's daily life. The news such as various bad events from natural phenomenal or climate is unpredictable. Sometimes fake news are broadcasted for an event that creates confusion among the people. So, there is a need of a system to detect it. Internet is one of the important tools which involves large number of persons as its user. There are different social media platforms which user used for [2]. Every user can make a post or spread the news through the online platforms. So, it is easy for the user to spread lies on Internet and they use fake accounts for it [10]. So there is growing demand for system which can predict the reliability of that news [3].

To summarize our survey, in our modern era where everyone relies on various online resources for news. Increasing social media platforms like Facebook, Twitter, etc. have facilitated the exponential increase in the number of fake news spread [6]. The spread of fake news has far-reaching consequences like from the creation of biased opinions to swaying election to violence between different societies for the benefit of certain candidates. This proposed model is made to differentiate between real and fake news and guide the people to believe on the true news.

In the following sections, proposed model and its features have been discussed in Sect. 3. The experimental analysis of various selected algorithms suitable for predicting label variable from the dataset has been discussed in detail along with its comparison graphs and reason for being selected in Sect. 4. The experimental result, the machine learning web application and its features have been discussed along with its interface design and working functionality after which why it is being called an automatic dataset updating application has been discussed with details and proofs

in Sect. 5. Finally, conclusion has been drawn from the experiments and researches and references have been given in Sect. 6.

3 Proposed Model

The proposed model will be a user-driven custom input web app which will predict the type of the news and the probability of its predicted type news on its own based on user input like the type of post whether it is a link or video or text, where it is posted, on which platform, etc. which is different from the existing models on Streamlit platform coded in Python by using random forest classifier algorithm to predict. Also the model can automatically increase its accuracy on its own without any human interaction by appending the user inputs with its original dataset on a real-time basis.

Figure 1 shows the workflow model of our proposed model where on giving the required inputs mentioned in the bubble boxes to the machine learning model will predict the type of news and show its prediction percentage.
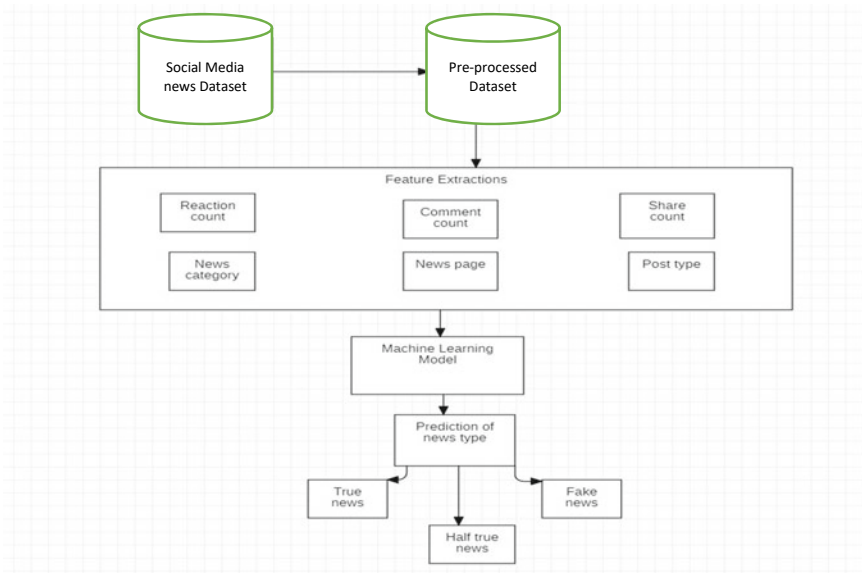

Fig. 1 Workflow diagram of proposed model

Category	Page	Post Type	Rating	Debate	share_cou	reaction_c	comment_	Day
2	1	4	4	0	96	146	15	19
2	1	4	4	0	25	207	108	19
3	8	1	1	0	52	131	25	19
3	8	1	3	0	1640	2676	284	19
3	8	1	1	0	748	1417	71	19
3	8	1	3	0	272	2409	53	19
3	8	1	1	0	309	2167	49	19
3	8	1	2	0	203	2115	40	19
3	8	1	2	0	33	204	28	19
3	8	1	1	0	87	707	24	19
3	8	1	2	0	244	813	77	19
3	8	1	3	0	51	86	12	19
1	9	1	3	0	483	5317	101	19
1	9	4	3	0	688	3329	37	19
3	4	1	3	0	14	35	9	19
1	9	1	3	0	4030	24066	331	19
2	1	1	3	0	2	37	25	19
1	9	1	3	0	33767	27185	3129	19
3	4	1	2	0	3304	1996	220	19
2	1	1	3	0	3	24	6	19
3	4	1	1	0	18	32	12	19
3	4	2	4	0	500	467	16	19
3	4	1	3	0	74	154	92	19

Fig. 2 First 23 values of the news dataset

4 Experimental Results and Analysis

4.1 Dataset

Figure 2 shows the modified dataset that we have taken for our research purpose from the raw dataset from Kaggle after converting all the categorical values to numerical values, filling missing values with the median set. This dataset comprises category of the post (whether it is a link or a page), page on which the post has been made, post type, debate on it, share count, reaction count, day of the post, and comment count as input columns and rating of the post as label column to be predicted by the app.

4.2 Correlation Metrix of Our Data

Figure 3 shows the correlation values between various columns of the dataset with given color indexes on the varying range of correlation values.

Since we are focused on predicting the type of the news, we take rating as the target variable.

We have taken three algorithms of classification for our dataset:

- Random Forest Classifier (RF).
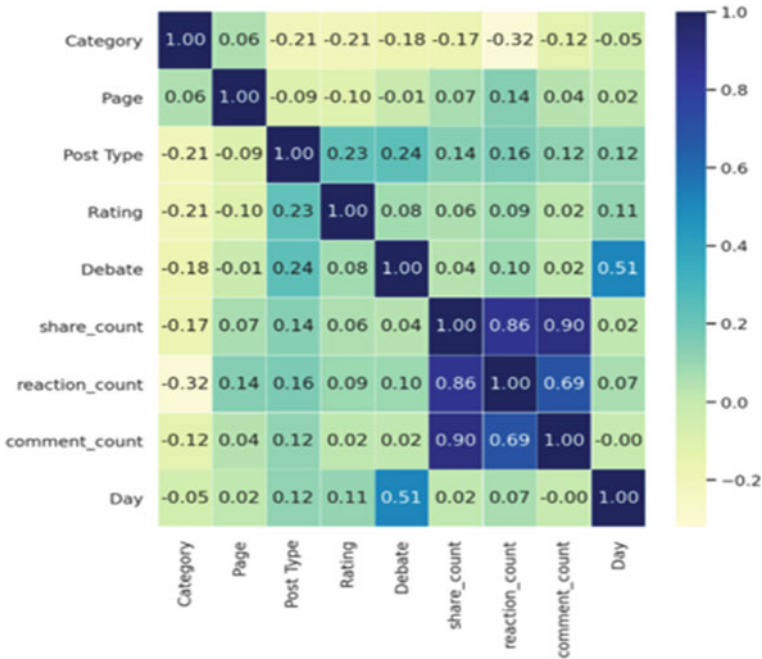- K-Nearest Neighbors Classifier (KNN).

Fig. 3 Correlation values between different variables in the dataset

- Logistic Regression Classifier (LR).

And we have done an analysis of f1-score, recall value, model score, accuracy, and ROC curve of the above algorithms on our dataset. We have also hyperparameterized each and every algorithm using normal parameter tuning and random search CV hyperparameter tuning.

4.3 Scores of Our Model

Table 1 depicts the model names used in this research analysis and its accuracy scores and it is clear from the table that RF comes to be the best with 78.1% score.

Table 1 Models and its accuracy scores

Model	Accuracy (%)
LR	73.96
KNN	70.45
RF	78.11
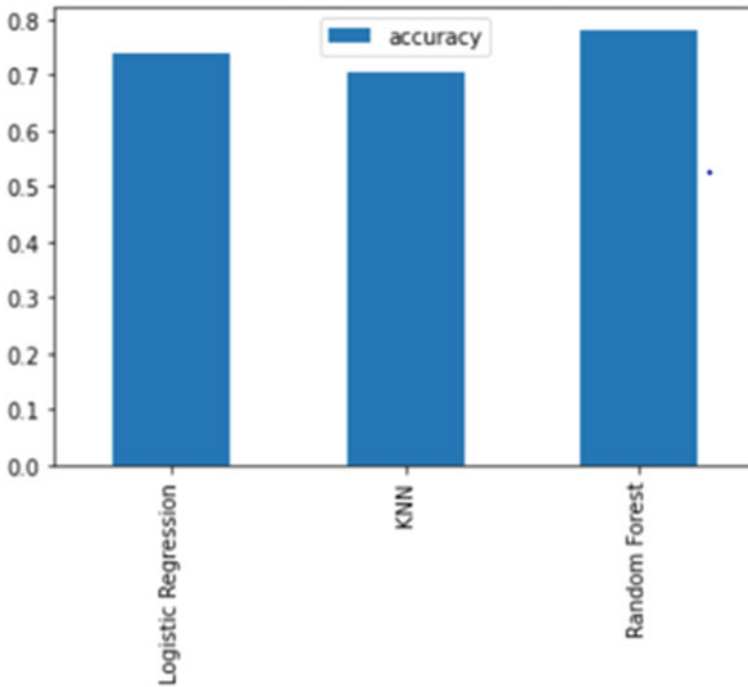
Fig. 4 Comparison graph between the model accuracy scores

4.3.1   Visual Representation of the Accuracy of the Models

Figure 4 depicts the graphical comparison and analysis between the accuracy scores of the three algorithms.

4.4   Score of KNN After Tuning

Figure 5 depicts the maximum accuracy score of KNN model after simple parameter tuning which comes to be approximately 75%.

4.5   Hyperparameter Tuning of RF and LR

Table 2 depicts the accuracy scores of logistic regression and random forest classifiers after hyperparameter tuning and KNN after simple parameter tuning. Since random forest accuracy scores come to be the best, it is being selected as the final model for predicting the news rating of the given dataset as shown in Fig. 1.
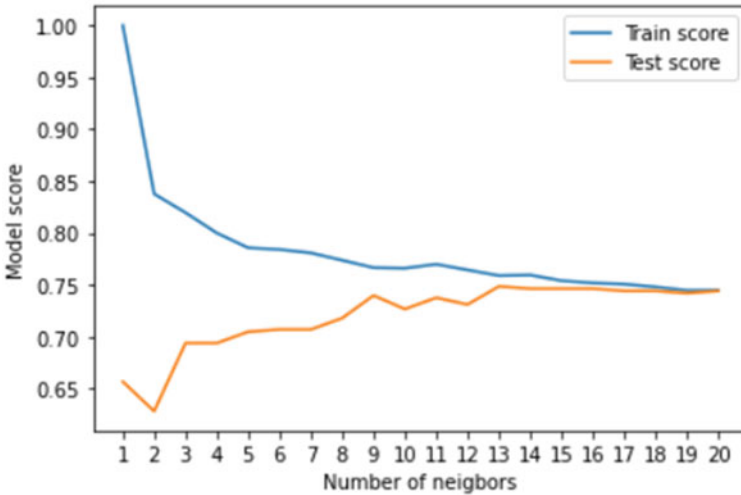
Maximum KNN score on the test data: 74.84%

Fig. 5 Train test scores of KNN after tuning

Table 2 Accuracy model scores of LR and RF after tuning

Model	Accuracy (%)
LR	**75.05**
RF	**79.43**

4.6 Classification Report

Table 3 depicts the classification report of our modified dataset with mixture of true and false, mostly false, mostly true, and no factual content being the rating of the news giving the precision (fraction of relevant values from retrieved data), recall (fraction values that are of interest from retrieved data), and f1-score (harmonic mean of precision and recall) values.

Table 3 Classification report of our dataset

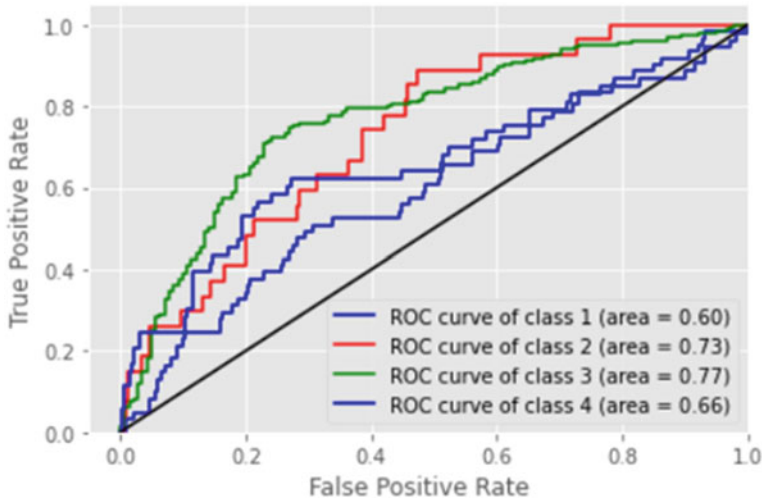| News type	Precision	Recall	F1-score
Mixture of true and false	0.43	0.21	0.29
Mostly false	0.67	0.11	0.18
Mostly true	0.82	0.96	0.88
No factual content	0.76	0.49	0.60
Macro-average	0.67	0.44	0.49
Weighted average	0.77	0.79	0.76

Fig. 6 ROC curve of the four classes, i.e., four types of news

4.7   ROC Curve

Figure 6 depicts the ROC curve of the different types of rating of the news. The index is mixture of true and false is class 1, mostly false type news is class 2, mostly true type news is class 3, and no factual content.

5   Fake News Predictor Application

5.1   Interface and Input

Figure 7 depicts the web app interface which will predict the rating of the news based on user input of the input columns mentioned in Fig. 1. This web app has been made using Streamlit library of Python 3.7 on which machine learning codes have been embedded with user interaction features of Streamlit. The entire dataset is the default training set whereas the user input feature is the test dataset.

Figure 8 depicts the prediction made by the fake news predictor app based on user input which comes to be mostly true. Also the app shows the prediction probability of all the rating types thus giving a proper justification of its own prediction.

Simple Fake News Prediction App

This app predicts the **News** type!

Input Fake News Data

Select the Category

left ▾

Select the Page

ABC News Politics ▾

Select the Post type

link ▾

Choose if debate happened or not

Yes ▾

Share Count

342925

1   1088995

Fig. 7 Predicting app features

Reaction Count

199325

2   456458

Comment Count

86086

1   159047

Day

21

19   27

User Input parameters

	Category	Page	Post Type	Debate	share_count	reaction_count
0	left	ABC News Politics	link	Yes	342925	199325

Make predictions

Prediction

The news type is: mostly true

Prediction Probability

	mixture of true and false	mostly false	mostly true	no factual content
0	0.1800	0.0400	0.5200	0.2600

Download Test_data.csv File

Fig. 8 Prediction made by the app on given features

5.2 Fake News Predictor App Automatic Updation of Dataset

The proposed model is a self-driven model who takes the user input along with its own dataset on its own, not involving any human interaction on its part. Hence, we can say the more the model is used by users across the globe, the model will increase its accuracy on its own by appending the user inputs also without any knowledge of the user. Figure 9 shows the strength of the dataset before prediction is made by the app and we can see that the total dataset strength is 2285.

Figure 10 shows the strength of the dataset after a prediction is made by the app which comes to be one more row than the previous figure, i.e., Fig. 9.

After making a prediction:

The total strength becomes 2286 and also the user input of the last try is captured thus becoming a live real-time artificial intelligence model. The working model can be found at https://share.streamlit.io/aritra1311/fake-news/predict.py.

2276	2	3	1	3	0	174	1394	626	27
2277	2	3	1	3	0	171	484	320	27
2278	3	4	1	3	0	108	162	5	27
2279	3	4	1	3	0	76	149	3	27
2280	2	3	1	3	0	66	543	163	27
2281	3	4	1	1	0	308	435	512	27
2282	3	4	1	3	0	100	173	10	27
2283	1	9	2	4	0	24499	47312	1375	27
2284	3	4	1	3	0	2	23	3	27
2285	2	3	1	3	0	8	95	47	27

Fig. 9 Number of entries before prediction

2277	2	3	1	3	0	171	484	320	27
2278	3	4	1	3	0	108	162	5	27
2279	3	4	1	3	0	76	149	3	27
2280	2	3	1	3	0	66	543	163	27
2281	3	4	1	1	0	308	435	512	27
2282	3	4	1	3	0	100	173	10	27
2283	1	9	2	4	0	24499	47312	1375	27
2284	3	4	1	3	0	2	23	3	27
2285	2	3	1	3	0	8	95	47	27
2286	1	1	1	3	1	342925	199325	86086	21

Fig. 10 Number of entries after prediction

6   Conclusion

Nowadays, we are very much dependent on social media for news but it is not sure that weather the news are real or fake. So, lot of times users are influenced by fake news and that leads to many problems like riots, violence, mob lynching, etc. So, to help the common people whether the news shared is fake or real news this web application can be very much useful. Here user can give some input related to the news and it will predict whether the news is real or fake or half-true. This machine learning model achieves an accuracy of 80% and this web application can help the user to be aware of fake news and aware others also. The automatic updation of dataset and data sharing without user involvement in this proposed model is an added advantage.

References

1. "Streamlit community". https://docs.streamlit.io/en/stable/ Streamlit official website" Accessed 10 June 2021
2. "AndrewGuess","Science Advance". https://advances.sciencemag.org/content/5/1/eaau4586?rss=1&fbclid=IwAR0AnmmBOuikMvGya9AUs3Zd0418CI4aeKLocjhhPfZIQJcgXbGyw3Ix-nE. Accessed 3 March 2021
3. Ahmed H. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. https://link.springer.com/chapter/10.1007/978-3-319-69155-8_9 Accessed 4 April 2021
4. Alim Al Ayub Ahmed (2021) Detecting fake news using machine learning: a systematic literature Review" https://arxiv.org/abs/2102.04458. Accessed 13 April 2021
5. Denis Zhuk. Methods to identify fake news in social media using artificial intelligence technologies. https://link.springer.com/chapter/10.1007/978-3-030-02843-5_36. Accessed 23 March 2021
6. Syed Ishfaq Manzoor (2021) Fake news detection using machine learning approaches: a systematic review. https://ieeexplore.ieee.org/abstract/document/8862770. Accessed 25 March 2021
7. Iftikhar Ahmad. Fake news detection using machine learning ensemble methods. https://www.hindawi.com/journals/complexity/2020/8885861/ Accessed 14 May 2021
8. Alvaro Figueria. The current state of fake news: challenges and opportunities. https://www.sciencedirect.com/science/article/pii/S1877050917323086. Accessed 1 June 2021
9. Denis Zhuk, Arsenii Tretiakov, Andrey Gordeichuk. Methods to identify fake news in social media using machine learning. https://fruct.org/publications/acm22/files/Zhu.pdf. Accessed 2 June 2021
10. Timothy Garton Ash. In the war of fake news versus facts, here's what the next battle should be. https://www.theguardian.com/commentisfree/2021/feb/08/fake-news-facts-facebook-twitter-fox-news-democracy. Accessed 14 June 2021