International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

Detecting Fake News using Machine Learning

Premjyoti Dhar[1], Shara Manger[2], Binod Subba[3], Sankalpa Rai[4], Susmita Rai[5], Shirshak Gurung[6]

[1-5]Student, Department of Computer Science and Technology, Centre for Computers and Communication Technology, Namchi, Sikkim, India

[6]Senior Lecturer, Department of Computer Science and Technology, Centre for Computers and Communication Technology, Namchi, Sikkim, India

---

Abstract - *In The 21st Century, due to excessive use of social media and internet, it has become an easy thing for people to receive news on a timely basis. This has led to the spread of fake news and hoaxes all over the internet which has proven to be a major threat for every other country in the world. The goal of this project was to create a model for classifying news into real or fake. For this we have used 4 Supervised Machine Learning Algorithms i.e, Logistic Regression, Decision Tree Classifier, Gradient Boosting Classifier and Random Forest Classifier in our model which predicts if the news is real or fake. The true and fake datasets used in our project have been collected from Kaggle. The obtained accuracy and confusion matrix results of each algorithm are also shown after the model is trained with the given dataset.*

***Key Words*: Classify, news, machine learning algorithms, social media, accuracy.**

1. INTRODUCTION

The term "Fake News" has become a very renowned word in most of our current lives and the major reason for this is the increasing number of fake news, information and rumors spread in online social medias and websites. To deal with the excessive spread of such hoaxes and news, people have tried to adapt various methods of crosschecking and blacklisting of fake authors and articles but then these methods haven't been able to yield consistent results. Thus, Machine Learning techniques come into play here as it can be used to find out the credibility of such news on the internet.

1.1 Data

There are two datasets namely "True.csv" and "Fake.csv", for our project that we have collected from Kaggle. The Fake dataset consists of 23503 rows of data from various news articles available on the internet whereas the True dataset consists of 21418 rows of data. The attributes of both the datasets are –

1. id – Unique ID for the news article.
2. title – Title of the news article.
3. text – The text of the news article. It might be incomplete in few cases.
4. subject – Type of news article.
5. date – Date of publication of the news article.

1.2 Supervised Learning

For our project we have used Supervised Machine Learning algorithms to classify our news data into true and fake. In supervised learning, the model is trained using labeled data where the output is also given to the model so that it can learn from that data and predict accordingly. We have used 4 Supervised ML Algorithms for our model which are as follows – 1) Logistic Regression 2) Decision Tree Classifier 3) Gradient Boosting Classifier 4) Random Forest Classifier.

1.3 Confusion Matrix

Confusion Matrix can be defined as a matrix which is used to evaluate the performance of a Machine Learning algorithm, usually a classification one.

The confusion matrix consists of the following values –

- True Positive (TP): It is a TP value if both the predicted and actual class is positive.
- False Positive (FP): It is a FP value if the predicted class is positive and the actual class was negative.
- False Negative (FN): It is a FN value if the predicted class is negative and the actual class was positive.
- True Negative (TN): It is a TN value if both the predicted and actual class is negative.

We can tell if a model is performing as expected if the Confusion Matrix result consists of true positive and true negative values.

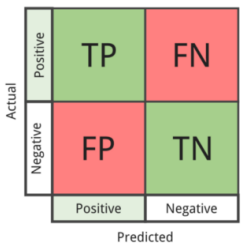

Fig-1: Representation of Confusion Matrix

2. METHODOLOGY AND MODELING

2.1 Methodology

1. At first, for manual testing, equal no. of data (in rows) from true and fake dataset is removed and merged into a single ".csv" file called manual_testing.csv.
2. Next the original true and fake datasets are merged into a single dataset.
3. Then, cleaning of the dataset is done by removing unnecessary columns, special characters, symbols and links from the original dataset.
4. Organizing of data is done using a function to convert text into lowercase.
5. Splitting of dataset is done into train and test split.
6. Vectorization of the dataset is done.
7. The working model consisting of 4 ML algorithms is trained using 75% of the original dataset.
8. The accuracy score and confusion matrix results of each algorithm used is calculated.
9. Finally the testing of the model is done by entering manual data from "manual_testing.csv" and the output is printed, which determines if the news is fake or not.

2.2 Software and Hardware Requirements

Table-1: Software and Hardware Requirements

Sl No.	Items	Specifications
1	Windows 10	Version 1703 and above
2	Jupyter Notebook	Version 3.0 and above
3	Pandas	Version 1.2.4
4	Numpy	Version 1.20.3
5	Scikit-learn	Version 0.24.2
6	Mlxtend	Version 0.18.0
7	Matplotlib	Version 3.4.2
8	String module	Version 0.1.2
9	RegEx module	Version 2021.7.6
10	Python	Version 3.8
11	Processor	2.3GHz
12	RAM	4GB and above

The working of the ML Model is represented using Flow Diagram, Context Flow Diagram(CFD) and Data Flow Diagram(DFD) respectively –

Flow Diagram

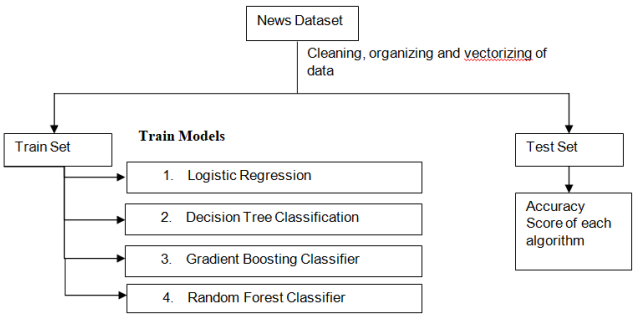

Fig-2: Flow Diagram of Working Model

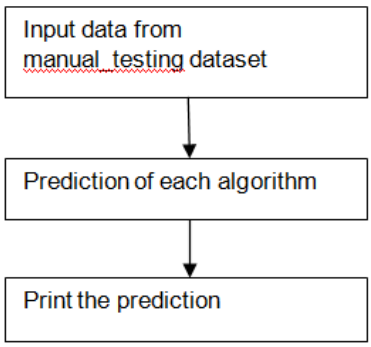

Fig-3: Flow Diagram of Manual Testing of Working Model

Context Flow Diagram(CFD)

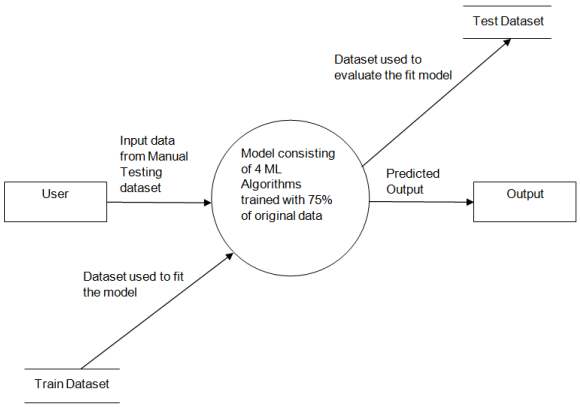

Fig-4: Context Flow Diagram

Data Flow Diagram(DFD)

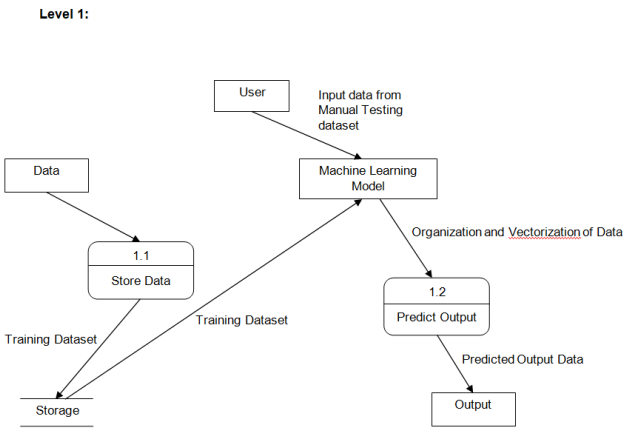

Fig-5: Level 1 Data Flow Diagram

3. RESULTS

The prediction of news done by our model is represented in the screenshot below:

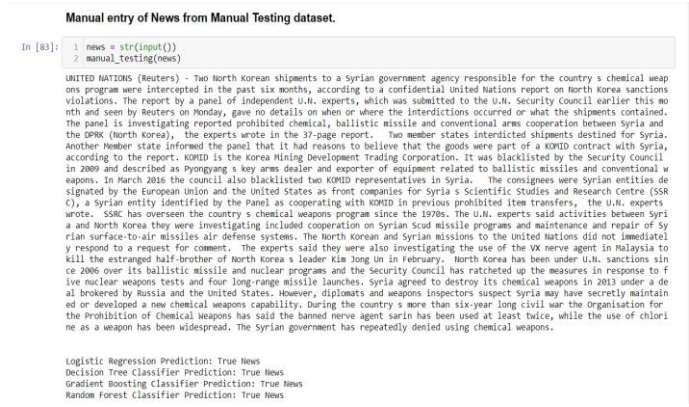

Fig-6: Screenshot from Jupyter Notebook representing the prediction done by the model

The accuracy and confusion matrix results of each algorithm used in our Machine Learning Model are illustrated in the table below:

Table 2: Comparison of all Algorithms used in the Machine Learning Model

Sl No.	Algorithms	Accuracy	True Positive	True Negative	False Positive	False Negative
1.	Logistic Regression	98.60	5801	5261	81	77
2.	Decision Tree Classifier	99.60	5857	5317	25	21
3.	Gradient Boosting Classifier	99.47	5835	5326	16	43
4.	Random Forest Classifier	98.83	5816	5273	69	62

4. CONCLUSIONS

The motive of developing the machine learning model is to differentiate fake news from true news. The datasets we have used in our work is collected from Kaggle, which contains both True and Fake news data, which is used to train our machine learning model for prediction of true or fake news. Our ML model was trained with sufficient amount of data so that it is able to obtain optimal accuracy. Some algorithms have achieved comparatively higher accuracy than others. We have used Confusion Matrix to compare the results of each algorithm.

At the end of the day, on being able to determine whether the news is real or fake can put an end to spreading of various kinds of hoaxes and rumors which are detrimental to the whole world.

ACKNOWLEDGEMENT

The successful completion of the Project was only possible due to the support we have received from our Parents and Teachers.

We express our gratitude to our HOD cum Project Coordinator Mr. Arvind Lal (Sr. Lecturer, Dept. of Computer Science and Technology) and Project Guide Mr. Shirshak Gurung (Sr. Lecturer, Dept. of Computer Science and Technology) for guiding us throughout the journey of the completion of the project.

REFERENCES

[1] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.

[2] H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383.

[3] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2018, pp. 272- 279.

[4] C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," 2017 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2017, pp. 208-215.

[5] Iftikhar Ahmad, Muhammad Yausaf, Suhail Yausaf and Muhhamma Dovais Ahmad, "Fake news detection learning essemble methods", Hindawi, vol. 2020, pages 1-11, October.

[6] Fathima Nada, Bariya Firdous Khan, Aroofa Maryam, Nooruz-Zuha, Zameer Ahmed "Fake news detection using logistic regression", Vol. 6 Issue 5, May 2019, IRJET.

[7] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 436-441.

[8] Shankar M. Patil, Dr. Praveen Kumar, "Data mining model for effective data analysis of higher education students using MapReduce" IJERMT, April 2017 (Volume-6, Issue-4).