iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42300

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Fight Misinformation and Detect Fake News Using Machine Learning Algorithm

Prof. Shweta Kahurke[1], Jyoti S.Haldar[2], Ashwini D. Bhagat[3], Rasika S. Andeo[4], Parag M. Dahikar[5], Ajinkya K.Kinhekar[6], Suraj S. Yelore[7]
[1]Assistant Professor, [2, 3, 4, 5, 6, 7]UG.Student, Department of Computer science and Engineering ,Shri Shankar Prasad Agnihotri Collage of Engineering, Wardha

Abstract: ***Machine learning and deep learning have been widely embraced, and even more widely misunderstood. In this article, I'd like to step back and explain both machine learning and deep learning in basic terms, discuss some of the most common machine learning algorithms, and explain how those algorithms relate to the other pieces of the puzzle of creating predictive models from historical data. As a discipline, machine learning explores the analysis and construction of algorithms that can learn from and make predictions on data.ML has proven valuable because it can solve problems at a speed and scale that cannot be duplicated by the human mind alone. With massive amounts of computational ability behind a single task or multiple specific tasks, mac Shines can be trained to identify patterns in and relationships between input data and automate routine processes***
Keywords: Random Forest algorithm, Logistic Regression, Long short-term Memory, ROC curves evaluation

I. INTRODUCTION

The fake news on social media and various other media is wide spreading and is a matter of serious concern due to its ability to cause a lot of social and national damage with destructive impacts. A lot of research is already focused on detecting it. This paper makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python scikit-learn, NLP for textual analysis. This process will result in feature extraction and vectorization; we propose using Python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer and Tiff Vectorizer. Then, we will perform feature selection methods, to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results.

II. LIBRARIES DETAIL

1) Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions like Active State's Active Python.

2) NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python.

3) Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

4) Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots

5) Natural Language Toolkit (NLTK) is a widely used, open-source Python library for NLP (NLTK Project, 2018). Several algorithms are available for text tokenization, stemming, stop word removal, classification, clustering, PoS tagging, parsing, and semantic reasoning. It also provides wrappers for other NLP libraries.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

III. WORK FLOW DIAGRAM

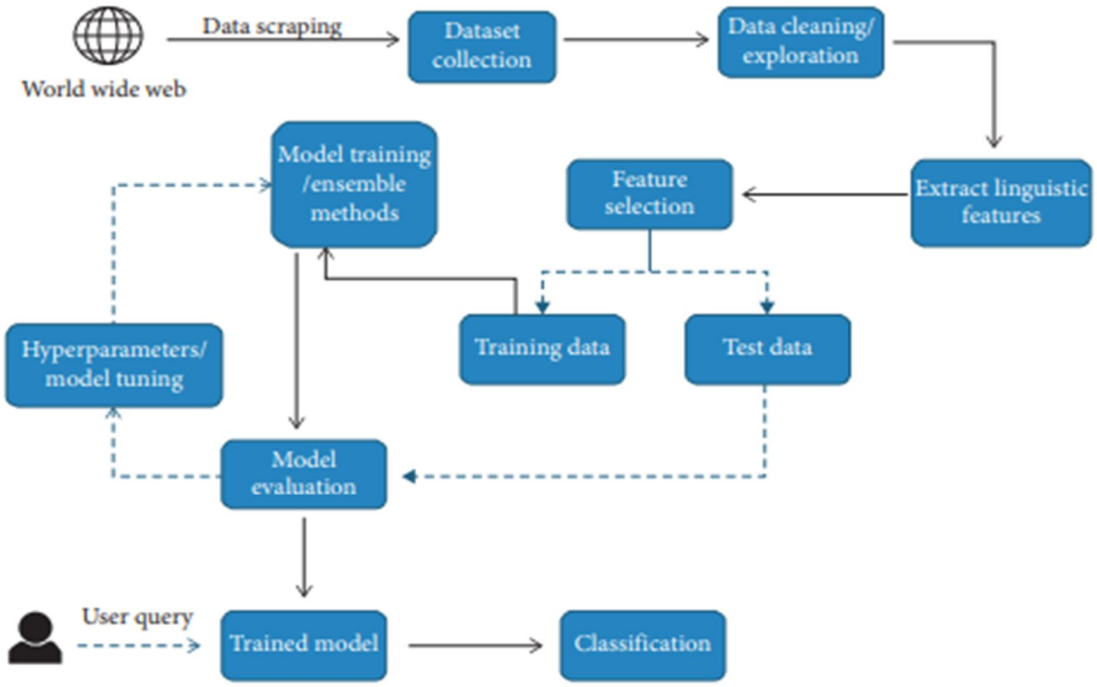
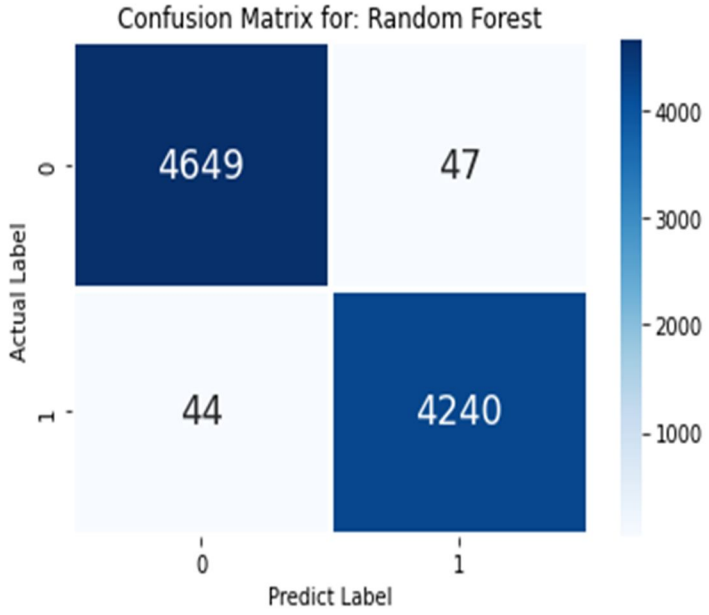


FIGURE 1: Workflow for training algorithms and classification of news articles.

IV. MACHINE LEARNING ALGORITHM DETAILS:

1) *Random Forest Algorithm:* Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

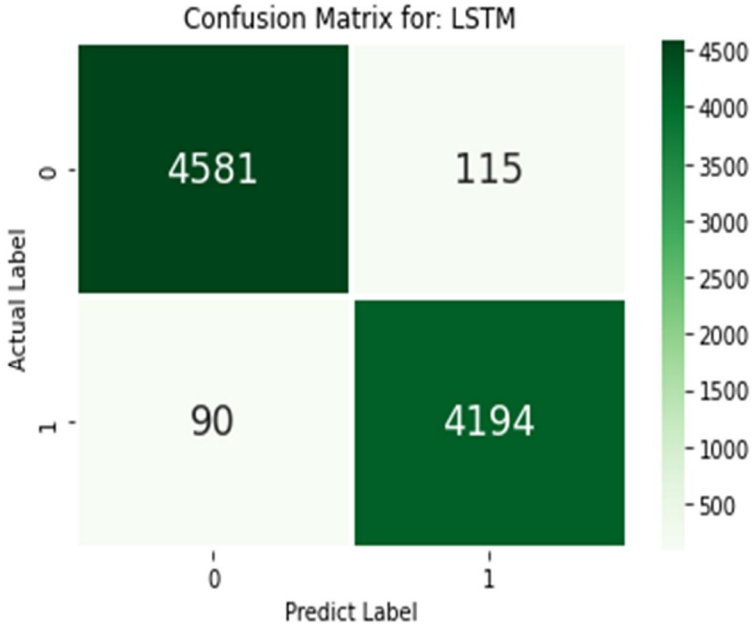


2) *Logistic Regression Algorithm:* Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for a given set of features (or inputs), X. Contrary to popular belief, logistic regression IS a regression model.

Mathematically, the logistic regression hypothesis function can be defined as follows [27]:

$h_\theta(X) =$

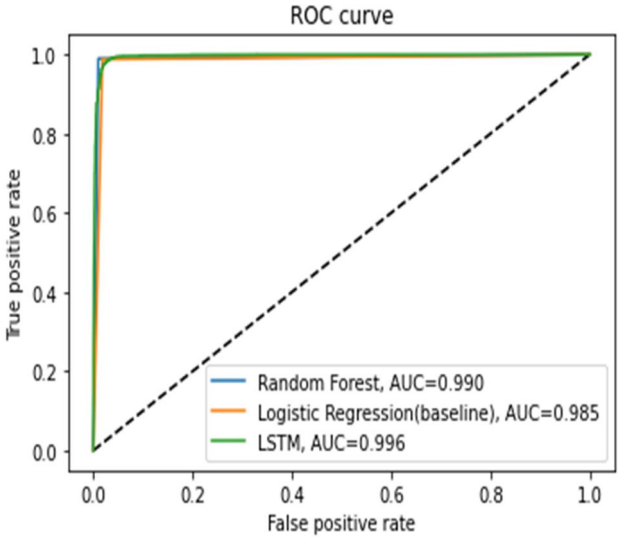
3) *Long Short-term Memory:* Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning (DL). Unlike standard feedforward neural networks, LSTM has feedback connections.

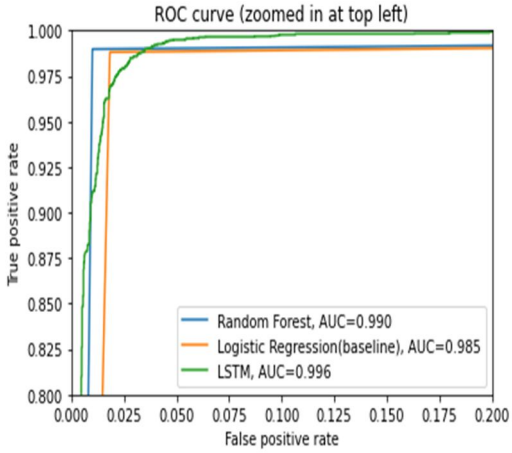


4) *Roc (Receiver Operating Characteristic Curve):* Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.

1) Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds.

2) ROC curves are appropriate when the observations are balanced between each class, whereas precision-recall curves are appropriate for imbalanced datasets.





V. DATASETS

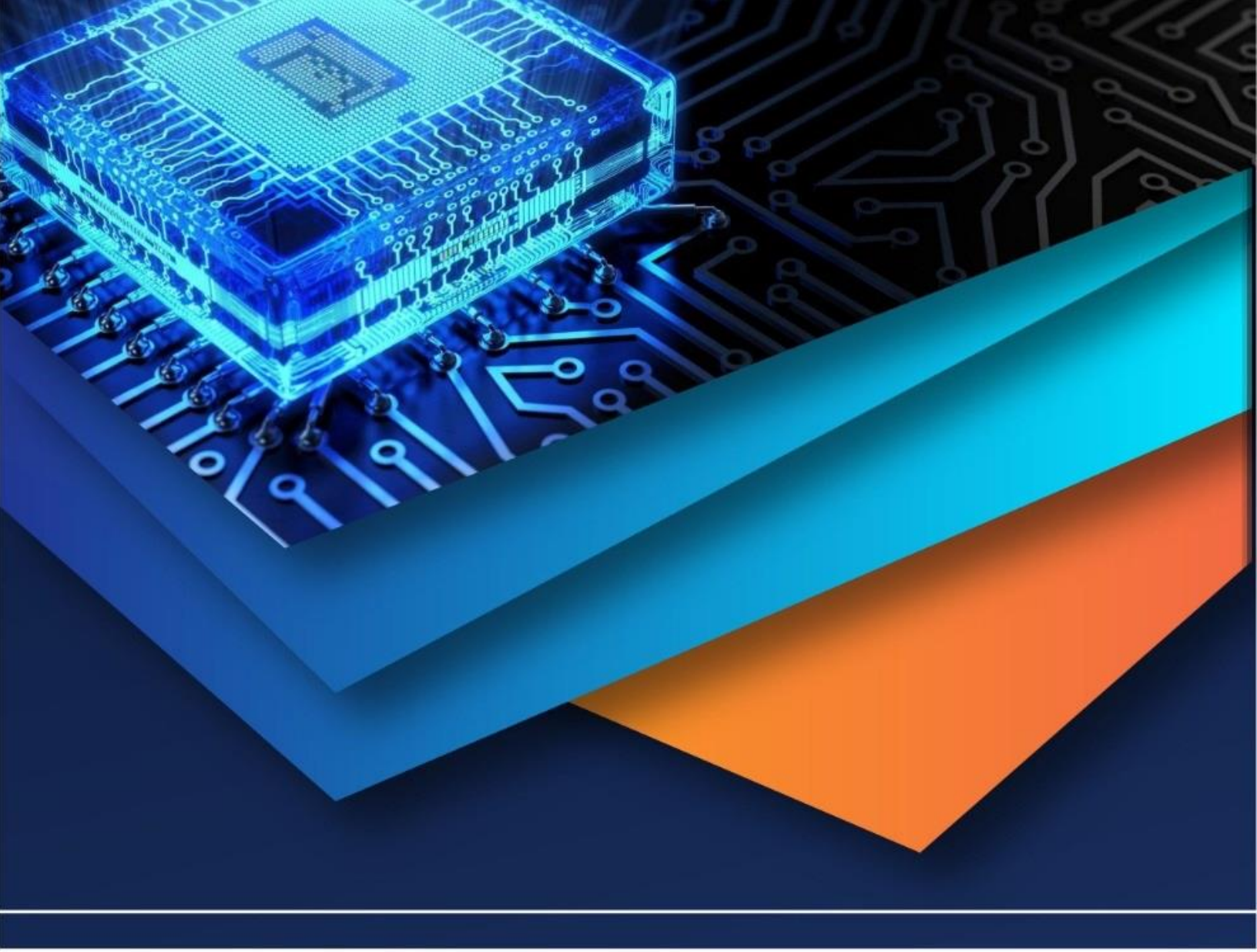
Datasets we used in this study are open source and freely available online. data includes both fake and truthful news articles from multiple domains. truthful news articles published contain true description of real-world events, while the fake news websites contain claims that are not aligned with facts. conformity of claims from the politics domain for many of those articles can be manually checked with fact checking websites such as politifact.com and snopes.com. We have used three different datasets in this study, a brief description of which is provided as follows. first dataset is called the "ISOT Fake News Dataset" which contains both true and fake articles extracted from the World Wide Web. true articles are extracted from reuters.com which is a renowned news website, while the fake articles were extracted from multiple sources, mostly websites which are flagged by politifact.com. dataset contains a total of 44,898 articles, out of which 21,417 are truthful articles and 23,481 fake articles. total corpora contain articles from different domains, but most prominently target political news.

VI. CONCLUSION

With the increasing popularity of social media, more and more people consume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which has strong negative impacts on individual users and broader society. In this article, we explored the fake news problem by reviewing existing literature in two phases: characterization and detection. In the characterization phase, we introduced the basic concepts and principles of fake news in both traditional media and social media. In the detection phase, we reviewed existing fake news detection approaches from a machine learning prospective, including feature of classification and regression by testing and training the data set model. We also further discussed the datasets, evaluation metrics, and promising future directions in fake news detection research and expand the field to other applications.

REFERENCES

[1] Kaggle, Fake News, Kaggle, San Francisco, CA, USA, 2018, https://www.kaggle.com/c/fake-news.
[2] Kaggle, Fake News Detection, Kaggle, San Francisco, CA, USA, 2018, https://www.kaggle.com/jruvika/fake-news-detection.
[3] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.
[4] T. M. Mitchell, 7e Discipline of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA, 2006.
[5] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY