Prediction of Fake News

Aruneswari S, CB.EN.U4ECE18107
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4ece18107@cb.students.amrita.edu

Bommineni Ramakrishna, CB.EN.U4ECE18110
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4ece18110@cb.students.amrita.edu

Chandhana S, CB.EN.U4ECE18112
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4ece18112@cb.students.amrita.edu

Nirali M Dave, CB.EN.U4ECE18136
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4ece18136@cb.students.amrita.edu

Prathiksha R R, CB.EN.U4ECE18144
Department of Electronics and Communication Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.u4ece18144@cb.students.amrita.edu

***Abstract*— It is difficult to distinguish between fake and real information due to the simple availability and exponential expansion of information available on publications available on the internet. The ease with which information can be shared has resulted in an exponential increase in its misrepresentation. When fraudulent information is widely disseminated, the articles' trustworthiness is likewise jeopardized. Thus, it has become a research challenge to automatically check the information viz a viz its source, content and publisher for categorizing it as false or true. Machine learning has played a vital role in classification of the information although with some limitations. This project reviews various Machine learning approaches in detection of fake news.**

Keywords—fake news, machine learning

I. INTRODUCTION

Traditionally people got news from various trusted sources that were required to keep up strict codes of practice. However, the presence of the internet has encouraged an advanced method to distribute, share data and news with almost no guidelines or article benchmarks. The Internet is an abundance of data and exceptionally worthwhile for different reasons. Due to the overwhelming information available on the internet, one must be cautious about the originality. Fake information is deliberately created and are purposefully or unexpectedly engendered over the internet. Creation and consumption of information over the internet have increased over time even if it's fake or real. Thus impacting groups of society who are large consumers of the internet and blinded by technology. Fake news is spread mainly for gaining political or financial incentives. They submit these well-crafted news stories and also recruit social bots or paid scammers to spread the news more rapidly and use different approaches using text-based analysis for detecting fake news.

This paper makes use of classification models in machine learning to detect if a news from the taken dataset is fake or real and compares the accuracy of 4 classification models – logistic regression, decision tree classification, gradient boosting classification and random forest classification.

Problem Statement

There has been a large surge of fake news in recent times due to the immense use of online news media**.** Hence it is extremely essential that certain measures should be taken in order to reduce or distinguish between real and fake news. So, we propose an ML model, and compare the accuracy of different classification algorithms.

II. RELATED WORK

The available literature has described many automatic detection techniques of fake news and deception posts. Since there are multidimensional aspects of fake news detection ranging from using chatbots for spread of misinformation to use of click baits for the rumor spreading. There are many click baits available in social media networks including Facebook which enhance sharing and liking of posts which in turn spreads falsified information. Lot of work has been done to detect falsified information.

- The Authors in [1] have described Linguistic Cue Approaches with Machine Learning, Bag of words approach, Rhetorical Structure and discourse analysis, Network analysis approaches and SVM classifiers. These are models are text based only and have very little or negligible improvement on existing methods.
- The authors of [2] have classified every tweet/post as binary classification Problem. The Classification is purely on the basis of source of the post/tweet. The

Authors used manually collected data sets using twitter API. The following algorithms were used on data sets

1. Naïve Bayes; 2. Decision trees; 3. SVM; 4. Neural Networks; 5. Random Forest; 6. XG Boost.

The results show 15 percent fake tweets, 45 % real tweets, rest posts were undecided.

- The authors of [3] have Introduced Need for hoax detection. They Used ML approach by combining news content and social content approaches. The authors Claim the performance is good as compared to described in literature. The authors Implemented it with Facebook messenger chatbot. Three different datasets of Italian news posts of Facebook where used. Both content-based methods with social and content signals using Boolean crowd sourcing algorithms where implemented. The following Methods were used by the: 1. Content based 2. Logistic regression on social signals. 3. Harmonic Boolean label crowdsourcing on social signals.
- [4] pointed out various sources of media and made the suitable studies whether the submitted article is reliable or fake. The paper utilizes models based on speech characteristics and predictive models that do not fit with the other current models.
- In [5] it has been discovered that fake news detection is a predictive analysis application. Detecting counterfeit messages involves the three stages of processing, feature extraction and classification. The hybrid classification model in this research is designed for Show fake news. The combination of classification is a combination of KNN and random forests. The execution of the suggested model is analysed for accuracy and recall. the final results improved by up to 8% using a mixed false message detection model.
- [6] used naïve Bayes classifier to detect fake news by Naive Bayes. This method was performed as a software framework and experimented it with various records from the Facebook, etc., resulting in an accuracy of 74%. The paper neglected the punctuation errors, resulting in poor accuracy.

III. Methodology

This section presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection phase, followed by preprocessing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers. Figure [1] describes the proposed system methodology. The methodology is based on conducting various experiments on dataset using the algorithms Logistic regression, Decision tree classification, Gradient Boosting classification and Random Forest. The experiments are to be conducted individually on each algorithm, for the purpose of achieving best accuracy and precision.

The main goal is to apply a set of classification algorithms to obtain a classification model in order to be used as a scanner for a fake news by details of news detection.

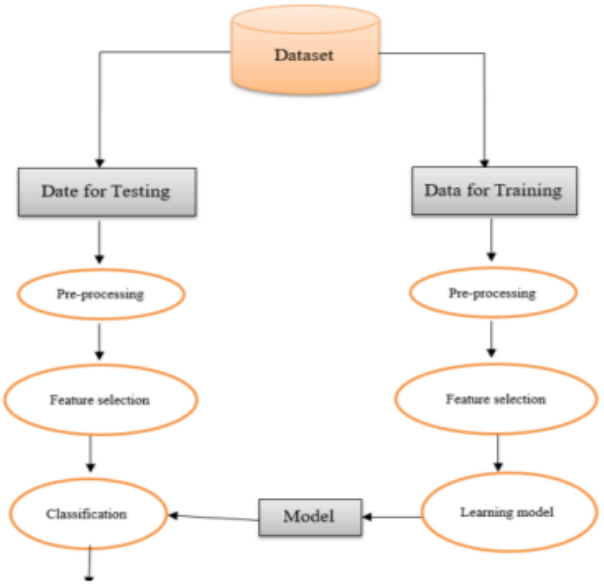


Fig. 1: Proposed System Methodology

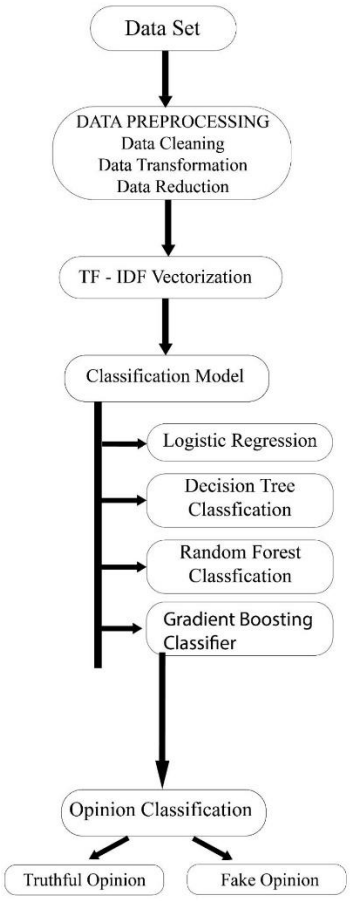


Fig. 2: Proposed Solution

IV. PROPOSED PROJECT

A. Overview of Proposed Solution

We selected the US news article dataset from the Kaggle website. This dataset is further split into Testing dataset and Training dataset.Upon collecting the data we are required to pre-process the data consisting of Data Cleaning, Data Removal and Data Transformation.The articles' unwanted variables such as authors, date posted, URL, and category are filtered out.The meaningless characters and duplicate characters are removed.These operations are performed on all the datasets to achieve consistency of format and structure.Once the relevant attributes are selected after the data cleaning and exploration phase, the next step involves extraction of the linguistic features. Linguistic features involved certain textual characteristics converted into a numerical form such that they can be used as an input for the training models.To accomplish the extraction of features from the corpus, we used TF-IDF Vectorization.The input features are then used to train the different machine learning models.The Classifier models used in this project are Logistic Regression Decision Tree classification, gradient boosting classification and random forest classification model. Each model is to be trained multiple times to optimize the model for the best outcome.The output of each model predicts whether the given news from the test dataset is Real or Fake News.To evaluate the performance of each model, we used accuracy and compared the accuracy of all the three models.

B. Vectorization

Machine Learning algorithms and almost all Deep Learning Architectures are not capable of processing strings or plain text in their raw form. In a broad sense, they require numerical numbers as inputs to perform any sort of task, such as classification, regression, clustering, etc. To build any model in machine learning, the final level data has to be in numerical form because models don't understand text or image data directly as humans do. So, vectorization is required to convert the text data into numerical vectors which are used to build machine models.

There are different methods of vectorization:

1. Count vectorization
2. Bag of words
3. N-grams Vectorization
4. TF-ID vectorization

a) TF-IDF vectorization: TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. As the frequency of the word appearing in the document increases, they are ranked lower, since they don't mean. The words that appear less frequently have a greater rank.

It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they don't mean much to that document in particular.

1. A histogram is created where every unique word is mapped to the occurrence of the word.
2. Term frequency: Term frequency denotes the frequency of a word in a document. For a specified word, it is defined as the ratio of the number of times a word appears in a document to the total number of words in the document.
   The expression is given by:

$$: \frac{Number\ of\ times\ \boldsymbol{term}\ appears\ in\ a\ document}{Total\ number\ of\ items\ in\ the\ document}$$

3. Inverse Document Frequency: It measures the importance of the word in the corpus. It measures how common a particular word is across all the documents in the corpus.
   It is the logarithmic ratio of no. of total documents to no. of a document with a particular word.
   The expression is given by:

$$\log\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ \boldsymbol{term}\ in\ it}\right)$$

4. The final numerical vector is calculated using: **Term Frequency * Inverse Documented Frequency**

C. Machine Learning models

a) Logistic Regression: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It provides a probability between 0 and 1, depending on which the output "TRUE" or "FALSE" is decided. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The logistic function is also called the sigmoid function. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

b) Decision Tree Classification: The decision tree is an important tool that works based on flow chart like structure that is mainly used for classification problems. Each internal node of the decision tree specifies a condition or a "test" on an attribute and the branching is done on the basis of the test conditions and result. Finally, the leaf node bears a class label that is obtained after computing all attributes. The distance from the root to leaf represents the classification rule. The amazing thing is that it can work with category and dependent variable. They are good in identifying the most important variables and they also depict the relation between the variables quite aptly. They are significant in creating new variables and features which is useful for data exploration and predicts the target variable quite efficiently.

Tree based learning algorithms are widely with predictive models using supervised learning methods to establish high accuracy. They are good in mapping non-linear relationships. They solve the classification or regression problems quite well and are also referred to as CART.

c) Random Forest Classification: Random forest classification is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. It contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

d) Gradient Boosting Classification:
Gradient boosting is a machine learning technique for regression, classification and other tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient boosted trees, which usually outperforms random forest. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

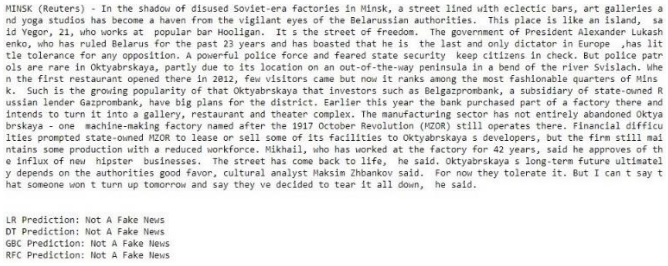
TOOLS USED

Python Jupyter, Machine Learning Models (Logistic Regression Decision Tree classification, gradient boosting classification and random forest classification model) and US news article Dataset.

V. RESULTS AND CONCLUSION

- Linear regression is a simple algorithm that can be implemented very easily to give satisfactory results. But it prone to overfitting and is sensitive to outliers.
- The Decision Tree is usually robust to outliers and can handle them automatically. This model is prone to overfitting problem.
- Gradient Boosting algorithm often provides predictive scores that are far better than other algorithms. Gradient Boosting Models will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting.
- Random Forest Classification is flexible to both classification and regression problems. It works well with both categorical and continuous values. It requires much computational power as well as resources as it builds numerous trees to combine their outputs.

Prediction of Fake and Real news:


MINSK (Reuters) - In the shadow of disused Soviet-era factories in Minsk, a street lined with eclectic bars, art galleries a
nd yoga studios has become a haven from the vigilant eyes of the Belarussian authorities. This place is like an island, sa
id Yegor, 21, who works at popular bar Hooligan. It s the street of freedom. The government of President Alexander Lukash
enko, who has ruled Belarus for the past 23 years and has boasted that he is the last and only dictator in Europe ,has lit
tle tolerance for any opposition. A powerful police force and feared state security keep citizens in check. But police patr
ols are rare in Oktyabrskaya, partly due to its location on an out-of-the-way peninsula in a bend of the river Svislach. Whe
n the first restaurant opened there in 2012, few visitors came but now it ranks among the most fashionable quarters of Mins
k. Such is the growing popularity of that Oktyabrskaya that investors such as Belgazprombank, a subsidiary of state-owned R
ussian lender Gazprombank, have big plans for the district. Earlier this year the bank purchased part of a factory there and
intends to turn it into a gallery, restaurant and theater complex. The manufacturing sector has not entirely abandoned Oktya
brskaya - one machine-making factory named after the 1917 October Revolution (MZOR) still operates there. Financial difficu
lties prompted state-owned MZOR to lease or sell some of its facilities to Oktyabrskaya s developers, but the firm still mai
ntains some production with a reduced workforce. Mikhail, who has worked at the factory for 42 years, said he approves of th
e influx of new hipster businesses. The street has come back to life, he said. Oktyabrskaya s long-term future ultimatel
y depends on the authorities good favor, cultural analyst Maksim Zhbankov said. For now they tolerate it. But I can t say t
hat someone won t turn up tomorrow and say they ve decided to tear it all down, he said.

LR Prediction: Not A Fake News
DT Prediction: Not A Fake News
GBC Prediction: Not A Fake News
RFC Prediction: Not A Fake News

Fig 3: Not a fake news


21st Century Wire says This week, the historic international Iranian Nuclear Deal was punctuated by a two-way prisoner swap
between Washington and Tehran, but it didn t end quite the way everyone expected. On the Iranian side, one of the U.S. citiz
ens who was detained in Iran, Nosratollah Khosravi-Roodsari, has stayed in Iran, but on the U.S. side all 7 of the Iranian
s held in U.S. prisons DID NOT show up to their flight to Geneva for the prisoner exchange with at least 3 electing to sta
y in the U.S TEHRAN SIDE: In Iran, 5 U.S. prisoners were released, with 4 of them making their way to Germany via Switzerla
nd.Will Robinson Daily MailNone of the Iranians freed in the prisoner swap have returned home and could still be in the Unit
ed States, it has been reported.The seven former inmates, who were released as part of a deal with the Islamic republic, did
not show up to get a flight to Geneva, Switzerland, where the exchange was set to take place on Sunday.Three of the Iranians
have decided to stay in the United States, ABC reported, with some moving in with their families. However it is not known wh
ere the other four are.Three of the Americans who had been detained in Iran Washington Post journalist Jason Rezaian, form
er U.S. Marine Amir Hekmati and Christian pastor Saeed Abedini left Tehran at around 7am the same day, but weren t met by
their counterparts in Switzerland Continue this story at the Mail OnlineREAD MORE IRAN NEWS AT: 21st Century Wire Iran Files

LR Prediction: Fake News
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News

Figure 4: Fake news

Performance metrics of different classifier models:

1. Logistic Regression:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5874
1	0.98	0.98	0.98	5346
accuracy			0.99	11220
macro avg	0.99	0.99	0.99	11220
weighted avg	0.99	0.99	0.99	11220

2. Decision Tree Classifier:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5874
1	1.00	1.00	1.00	5346
accuracy			1.00	11220
macro avg	1.00	1.00	1.00	11220
weighted avg	1.00	1.00	1.00	11220

3. Gradient Boosting Classifier:

	precision	recall	f1-score	support
0	1.00	0.99	1.00	5874
1	0.99	1.00	1.00	5346
accuracy			1.00	11220
macro avg	1.00	1.00	1.00	11220
weighted avg	1.00	1.00	1.00	11220

4. Random Forest Classifier:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5874
1	0.99	0.99	0.99	5346
accuracy			0.99	11220
macro avg	0.99	0.99	0.99	11220
weighted avg	0.99	0.99	0.99	11220

From the observed performance metrics,

- Fake news detection on the given dataset is performed by making use of Logistic Regression, Decision Tree Classification, Random Forest Classification and Gradient Boosting Classification.

- Performance metrics of 4 different classifier models are observed.

- While the accuracy of all four models is either 0.99 or 1, decision tree classifier model and gradient boosting classifier model predict if a news is fake or real with the highest accuracy.

VI. REFERENCES

[1] Conroy, N. J., Rubin, V. L., & Chen, Y. (2015, November). Automatic deception detection: Methods for finding fake news. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (p. 82). American Society for Information Science.

[2] Helmstetter, S., & Paulheim, H. (2018, August). Weakly supervised learning for fake news detection on Twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 274-277). IEEE.

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

[4] MykhailoGranik and VolodymyrMesyura. “Fake news detection using naive Bayes classifier.” First Ukraine Conference on Electrical and Computer Engineering (UKRCON). Ukraine: IEEE. 2017.

[5] Looijenga, M. S. “The Detection of Fake Messages using Machine Learning.” 29 Twente Student Conference on IT, Jun. 6th, 2018, Enschede, The Netherlands. Netherlands: essay.utwente.nl. 2018.

[6] Gilda, S. “Evaluating machine learning algorithms for fake news detection.” 15th Student Conference on Research and Development (SCOReD) (pp. 110-115). IEEE. 2017.