IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CLASSIFYING FAKE NEWS ARTICLES USING MACHINE LEARNING

[1]Lalitha Siripurapu, [2]Ramarapu Bangari, [3] D.Avinash Babu,

[1]Under Graduate Student, [2]M.Tech, Assistant Professor, [3]M.Tech, Assistant Professor
[1]Department Of Computer Science And Engineering,
[1]Satya Institute Of Technology And Management, Vizianagaram, India

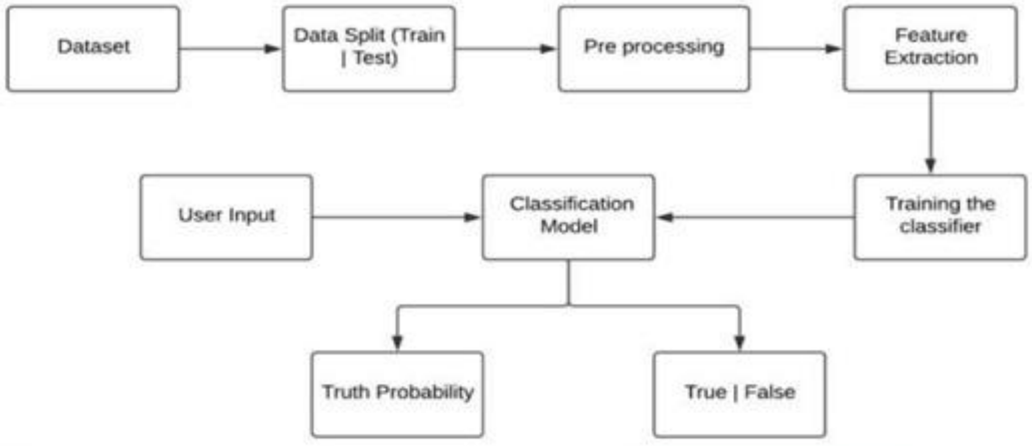
Abstract: A number of linguistic traits are common to the fake news reports that are first spread over social media platforms, including an overuse of unfounded hyperbole and unattributed cited information. The performance of a fake news classifier is documented in a study on fake news identification, the results of which are provided in this publication. A innovative fake news detector that leverages quoted attribution in a Bayesian machine learning system as a major feature to evaluate the likelihood that a news story is fraudulent was developed using the Text blob, Natural Language, and SciPy Toolkits. The method precision as a result is 63.333% effective in determining the probability of forgery in an article containing quotes, Influence mining is the term for this procedure, which is described as a unique tool that can be used to detect propaganda and even fake news.

- *Index Terms* - **Natural Language**,**Fake News**,**Fake NewsStories**,**Machine Learning**,**Social Media**,**Classification Performance**,**Twitter**,**Social Media Platforms**,**Integration Problems**,**Radio News**,**Research Team**,**Training Set**, **News Quality**

I. INTRODUCTION

Deliberately false content distributed as real journalism—also called "fake news"—is a global problem with information accuracy and integrity that affects how people vote, form opinions, and make decisions. Most fake news originates on social media platforms like Facebook and Twitter and then finds its way onto channels that are used by traditional media, such as radio and television news. The false news pieces that initially circulate on social media platforms share a number of language characteristics, such as an excessive use of unjustified hyperbole and unattributed quoted content. The results of an experiment into fake news identification, which describes the efficacy of a fake news classifier, are presented and discussed in this work.

II. METHODOLOGY



2.1 Static Search Implementation-

In static part, we have trained and used 3 out of 4 algorithms for classification. They are Naïve Bayes, Random Forest
and Logistic Regression.

Step 1: In first step, extracting features from the already pre-processed dataset.
Step 2: Here, built all the classifiers for predicting the fake news detection. The extracted features are fed into different classifiers.
Step 3: Once fitting the model, comparing the f1 score and checked the confusion matrix.
Step 4: After fitting all the classifiers, 2 best performing models were selected as candidate models for fake news classification.
Step 5: Now performed parameter tuning by implementing Grid Search CV methods on these candidate models and chosen best performing parameters for these classifier.
Step 6: Finally selected model was used for fake news detection with the probability of truth.
Step 7: Our finally selected and best performing classifier was Logistic Regression which was then saved on disk. It will be used to classify the fake news.

2.2 Dynamic Search Implementation-

In this search field we have used Natural Language Processing for the first search field to come up with a proper solution for the problem, Our application uses NLP techniques like Count Vectorization and TF-IDF Vectorization before passing it through a Passive Aggressive Classifier to output the authenticity as a percentage probability of an article.

Working-

The problem can be broken down into 3 statements-
1. Use NLP to check the authenticity of a news article.
2. If the user has a query about the authenticity of a search query then we he/she can directly search on our platform and using our custom algorithm we output a confidence score.
3. Check the authenticity of a news source.

III. Technology

NATURAL LANGUAGE PROCESSING

NLP stands for Natural Language Processing. It is the branch of Artificial Intelligence that gives the ability to understand and process human languages. Human languages can be in the form of text or audio format.

NLP is used in a wide range of applications, including machine translation, sentiment analysis, speech recognition, chatbots, and text classification. Some common techniques used in NLP include:

1. Tokenization: the process of breaking text into individual words or phrases.
2. Part-of-speech tagging: the process of labeling each word in a sentence with its grammatical part of speech.
3. Named entity recognition: the process of identifying and categorizing named entities, such as people, places, and organizations, in text.
4. Sentiment analysis: the process of determining the sentiment of a piece of text, such as whether it is positive, negative, or neutral.
5. Machine translation: the process of automatically translating text from one language to another.
6. Text classification: the process of categorizing text into predefined categories or topics.

Natural Language Toolkit (NLTK)

NLTK is python’s API library for performing an array of tasks in human language. It can perform a variety of operations on textual data, such as classification, tokenization, stemming, tagging, Leparsing, semantic reasoning etc.

Installation:

NLTK can be installed simply using pip or by running the following code.

```
! pip install nltk
```

or

```
Import nltk
nltk.download(‘all’)
```

IV.Algorithm

We employed a variety of algorithms during the feature extraction process, selecting the most accurate one.

4.1 Logistic regression

it is a supervised machine learning algorithm used for **classification tasks** where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors. The article explores the fundamentals of logistic regression, it’s types and implementations.

For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It’s referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

Logistic regression model accuracy (in %): **95.6140350877199**

4.2 NAIVE BAYES

1. Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
2. It is mainly used in **text classification** that includes a high-dimensional training dataset.
3. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
4. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
5. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Bayes' Theorem:

1. Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
2. The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

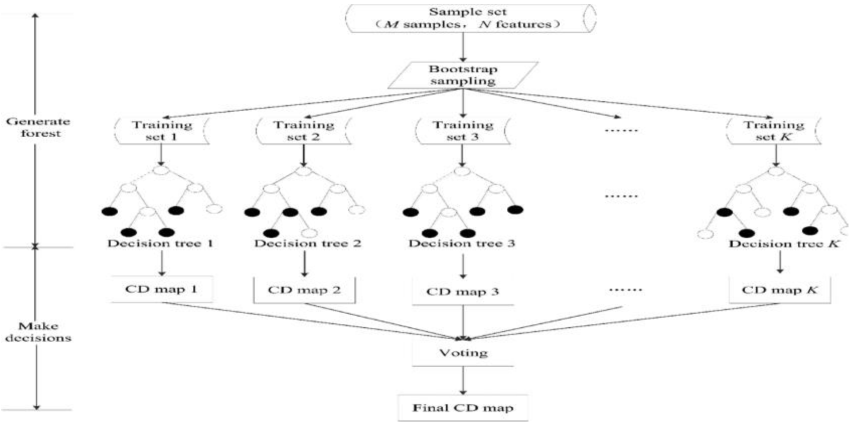
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

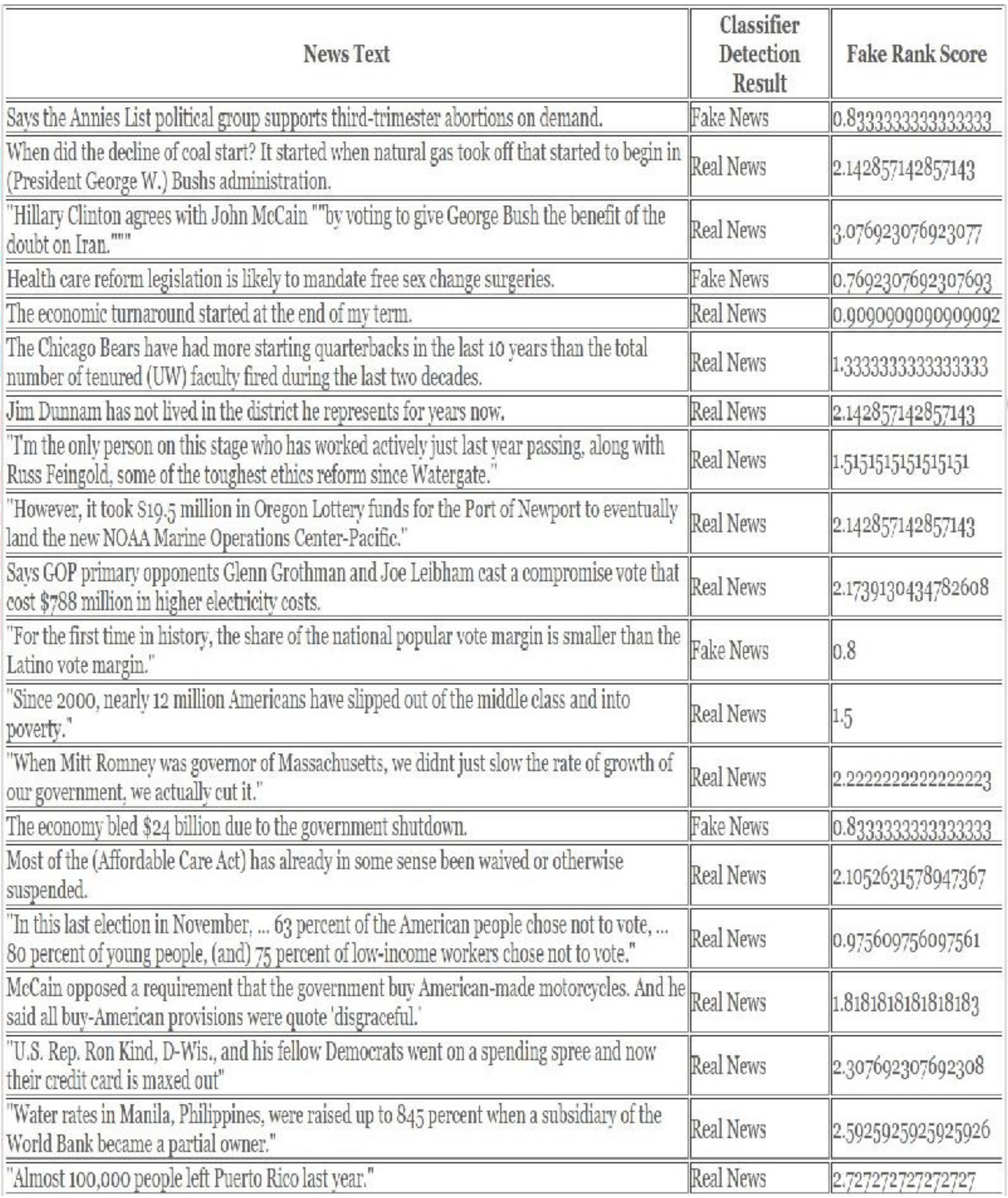
P(B) is Marginal Probability: Probability of Evidence.

4.3 RANDOM FOREST

Random Forest algorithm is a powerful tree learning technique in **Machine Learning**. It works by creating a number of **Decision Tress**. during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of **Overfitting** and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative decision-making process, supported by multiple trees with their insights, provides an example stable and precise results. Random forests are widely used for classification and regression functions, which are known for their
ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments.



V.RESULTS AND DISCUSSION



News Text	Classifier Detection Result	Fake Rank Score
Says the Annies List political group supports third-trimester abortions on demand.	Fake News	0.8333333333333333
When did the decline of coal start? It started when natural gas took off that started to begin in (President George W.) Bushs administration.	Real News	2.142857142857143
"Hillary Clinton agrees with John McCain ""by voting to give George Bush the benefit of the doubt on Iran."""	Real News	3.076923076923077
Health care reform legislation is likely to mandate free sex change surgeries.	Fake News	0.7692307692307693
The economic turnaround started at the end of my term.	Real News	0.9090909090909092
The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW) faculty fired during the last two decades.	Real News	1.3333333333333333
Jim Dunnam has not lived in the district he represents for years now.	Real News	2.142857142857143
"I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate."	Real News	1.5151515151515151
"However, it took $19.5 million in Oregon Lottery funds for the Port of Newport to eventually land the new NOAA Marine Operations Center-Pacific."	Real News	2.142857142857143
Says GOP primary opponents Glenn Grothman and Joe Leibham cast a compromise vote that cost $788 million in higher electricity costs.	Real News	2.1739130434782608
"For the first time in history, the share of the national popular vote margin is smaller than the Latino vote margin."	Fake News	0.8
"Since 2000, nearly 12 million Americans have slipped out of the middle class and into poverty."	Real News	1.5
"When Mitt Romney was governor of Massachusetts, we didnt just slow the rate of growth of our government, we actually cut it."	Real News	2.2222222222222223
The economy bled $24 billion due to the government shutdown.	Fake News	0.8333333333333333
Most of the (Affordable Care Act) has already in some sense been waived or otherwise suspended.	Real News	2.1052631578947367
"In this last election in November, ... 63 percent of the American people chose not to vote, ... 80 percent of young people, (and) 75 percent of low-income workers chose not to vote."	Real News	0.975609756097561
McCain opposed a requirement that the government buy American-made motorcycles. And he said all buy-American provisions were quote 'disgraceful.'	Real News	1.8181818181818183
"U.S. Rep. Ron Kind, D-Wis., and his fellow Democrats went on a spending spree and now their credit card is maxed out"	Real News	2.307692307692308
"Water rates in Manila, Philippines, were raised up to 845 percent when a subsidiary of the World Bank became a partial owner."	Real News	2.5925925925925926
"Almost 100,000 people left Puerto Rico last year."	Real News	2.727272727272727

VI.References

1.M. Balmas, "When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy Alienation and Cynicism", Communic. Res., vol. 41, no. 3, pp. 430-454, 2014.

Show in Context CrossRef Google Scholar

2.C. Silverman and J. Singer-Vine, "Most Americans Who See Fake News Believe It New Survey Says" in BuzzFeed News, Dec 2016.

Show in Context Google Scholar .

3.P. R. Brewer, D. G. Young and M. Morreale, "The Impact of Real News about '"Fake News"': Intertextual Processes and Political Satire", Int. J. Public Opin. Res., vol. 25, no. 3, 2013.

Show in Context CrossRef Google Scholar

4.D. Berkowitz and D. A. Schwartz, "Miley CNN and The Onion", Journal. Pract., vol. 10, no. 1, pp. 1-17, Jan. 2016.

Show in Context Google Scholar

5.C. Kang, "Fake News Onslaught Targets Pizzeria as Nest of Child-Trafficking" in New York Times, Nov 2016.

Show in Context Google Scholar

6.C. Kang and A. Goldman, "In Washington Pizzeria Attack Fake News Brought Real Guns" in New York Times, Dec 2016.

Show in Context Google Scholar

7.R. Marchi, "With Facebook Blogs and Fake News Teens Reject Journalistic "Objectivity"", J. Commun. Inq., vol. 36, no. 3, pp. 246-262, 2012.

Show in Context CrossRef Google Scholar

8.C. Domonoske, "Students Have 'Dismaying' Inability o Tell Fake News From Real Study Finds" in Natl. Public Radio Two-w., 2016.

Show in Context Google Scholar

9.H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election", J. Econ. Perspect., vol. 31, no. 2, 2017.

CrossRef Google Scholar

10.C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini and F. Menczer, The spread of fake news by social bots.

Google Scholar

11.A. Gupta, H. Lamba, P. Kumaraguru and A. Joshi, "Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy" in WWW 2013 Companion, 2013.

CrossRef Google Scholar

12.E. Mustafaraj and P. T. Metaxas, The Fake News Spreading Plague: Was it Preventable?.

Google Scholar

13.M. Farajtabar et al., Fake News Mitigation via Point Process Based Intervention.

Google Scholar

14.M. Haigh, T. Haigh and N. I. Kozak, "Stopping Fake News", Journal. Stud., vol. 19, no. 14, pp. 2062-2087, Oct. 2018.

CrossRef Google Scholar

15.O. Batchelor, "Getting out the truth: the role of libraries in the fight against fake news", Ref. Serv. Rev., vol. 45, no. 2, pp. 143-148, Jun. 2017.

CrossRef Google Scholar