



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 Issue: X Month of publication: October 2021

DOI: <https://doi.org/10.22214/ijraset.2021.38609>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Fake News Detection Using Various Machine Learning Algorithms

Gummadi Rohith Madhu Chandra¹, Sri Harsha Samayamanthula², Gogineni Hrushikesh Madhav³, Syed Afzal Ahammed⁴

^{1, 2, 3, 4}School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

Abstract: Reports have been around for a long time and have a track record of producing accurate results. The fast development of online media stages has increased the adverse consequence of bits of gossip; it accordingly gets essential to early identify them. Numerous techniques have been acquainted with distinguishing reports utilizing the substance or the social setting of information. In any case, most existing strategies overlook or don't investigate viably the engendering example of information in online media, including the grouping of collaborations of web-based media clients with news across time. In this paper, we present a new conversation identification method based on deep learning. By understanding the customers' depiction and the worldwide interconnection of clients' responses, our method makes advantage of the news dissemination dynamics. We target furnishing clients with a stage to check the bits of gossip they hear and counterfeit news that courses through online media stages like Twitter and Facebook. The clients can gather validated realities and check for the wellsprings of the news occasions continuously. Official groups can use the constant stream elements to follow an occasion and individuals engaged with the stream. Tests conducted on Twitter and Weibo datasets indicate that the suggested approach is cutting-edge in terms of performance.

Keywords: Propagation Dynamics, User Representation, Data Sets, Counterfeit

I. INTRODUCTION

Reports are things of unsubstantiated coursing data, which have been known for genuine outcomes. The development of online media stages makes ripe ground for bits of hearsay, consequently delivering talk recognition of extraordinary importance. Nonetheless, recognizing reports is a difficult errand; contemplates have revealed that people are bad at recognizing tales. Then again, analysts have examined bits of gossip according to various perspectives. There exist two conspicuous methodologies for talk identification: substance-based and social-setting-based methodologies. In the substance-based methodology, bits of gossip are distinguished depending on the substance of information and earlier information removed from huge information sources or the composing style of the news. On the other hand, the social context-based approach utilizes the social commitment of online media users such as Twitter responses. Using this technique, a large number of customer evaluations may be tallied, revealing the news's degree of veracity. Furthermore, methods based on social settings may reveal the news's hidden transitory spread example. Thus, the social-setting-based methodology has as of late become well known because of its great exhibition and the accessibility of extra data. In this work, we address the issue of talk discovery via online media utilizing social setting data. We see it as a two-class parallel characterisation problem, with non-gossip and conversation as the two classes. We discovered a few features in the proliferation cycle of information via friendly media clients by analysing existing datasets, such as the Twitter and Weibo datasets. Initially, there is a difference in the number of postings devoted to reports and genuine news throughout time periods. Besides, a few clients are more powerless against misdirecting data than others. Thus, these clients will in general be associated with the spreading of numerous reports in web-based media. Enlivened by these perceptions, we expect to identify bits of hearsay by perceiving the idiosyncrasies of the proliferation interaction of the news. To that aim, we propose Dual RNN for Rumor Recognition, a new engendering-driven model for discourse discovery that relies on intermittent neural networks (RNNs) (DRRD). Researchers at the University of Michigan released a report on lying in India on April 18, 2020, especially after Prime Minister Narendra Modi announced the referendum deadline on March 22, 2020. Lock down after two days to prevent the spread of the virus. COVID-19. [1] According to the study, the number of false detection events increased from two in the third period of January 2020 to 60 in the main seven-day period in April 2020. Despite the fact that false tales of a COVID-19 cure decreased over this time, the study showed that false instances that affected people actually increased. The investigation used 243 unique deception instances from a Tattle Civic Innovation whitepaper (a Delhi-based news project that means to make exact data more available to versatile first clients). Between January 23 and April 12, 2020, six fact-checkers – AltNews, BOOMlive, Factly, IndiaToday Fact Check, Quint Webqoof, and NewsMobile Fact Checker – revealed stories that were verified by the International Fact-Checkers Network (IFCN). [2]

II. PROPOSED METHODOLOGY

A. Propagation Dynamics

To take advantage of the web-based media's spread dispersion pattern, the appropriate social postings must be coordinated following a sequential request, i.e. via partitioning. Divide publications into sections for different time periods, for example, until the amounts of publications in the extensions are the same. However, we argue that the splitting method ignores internal fluctuations in the number of relays in the newscasting cycle. As a result, we use a common dividing technique of collecting postings by the hour.

Time stamps on recent posts on a specific event, such as the most recent one, indicate when that occasion first appeared. Furthermore, the hour list of the post makes a difference between the important post and the most recent post in terms of when it was posted. Each occasion's posts are then grouped together in the same division. An arrangement of hour segments addresses an event appropriately. We present an exceptional cushioning and scaling strategy to advance the variety of posts in allotments.[3]

1) *Formulation and Notation of the Problem:* To solve the rumor detection problem we use social context information. Suppose a news item covers an event and $E = \{e_i | N_i = 1\}$ do this as a set of events. Suppose SI represents a set of social interactions related to an event e_i then $SI = \{(p_j, u_j, t_j) | M\}$, where p_j refers to the social post, u_j refers to the person posting the post and t_j is the associated timestamp. Let N be the number of iterations and let L be the set of its binary labels. Our goal is to develop a mathematical model F , which can predict the probability of an event becoming a rumor based on its SI interactions SI ie $P(e_i = 1 | SI) = F(SI)$. We study the set of social interactions Deadline T $STs = \{(p_j, u_j, t_j) | t_j \leq T\}$ defines the set of social interactions I formed before the expiration date I , then rumors I probability from event I within T ($e_i = 1 | ST$) = $F(ST)$.

2) *Model Intuition and Structure:* Our model is built on recurrent neural networks, as shown in Fig. 1. Text, User, and Integration are the three components that make up the system.

a) *The Text Module:* Some studies have shown that during certain time periods, the frequency of inquiry terms is much greater among rumour postings than among non-rumor posts. There is also a difference in the amount of social postings about rumours vs real news, as seen in Fig 1. If you're looking for specific examples, go no further than the content module. A good place to start is by training the doc2vec model on a corpus of social media postings about the events in the training set. This method has proven useful in many NLP-related tasks. Every friendly post receives an implanted D-V measurement as a result of using the pre-prepared doc2vec model As a result, the posts' embeddings around the same time period come to the midway of component knowledge, creating the segment's picture. We utilize personality vectors, i.e., vectors with every one of the 1 passages to address parcels that contain no posts. As a result, an event is organized by a framework $X R^{nd} v$ where n represents the number of clock components. The number of posts in the K -th segment is measured by a logarithmic coefficient $c_k = \log(m_k + 1) + 1$, where m_k is the number of posts in the k -th segment. The purpose of this measurement is to calculate a wide range of message numbers between packets. The M logarithm is also used to smooth the coefficients, because the height of m_k varies greatly between sections; For example, in the Weibo dataset, the number of relays per hour ranges from 1 to 24192. Next, the expanded and scaled representation is presented into a two-layer RNN. In comparison to the long transient memory partner (LSTMs) that was delivered in [26], we chose the gated repetitive units (GRUs) engineering because it is easier to prepare. At this point we use max-pooling-over-the-long-hall to get the highlight vector XFR_{df} for the text unit $hn_{R_{df}}$ represents the infinite steps $k = 1, \dots, n$ df return vector element from RNN. To be more precise, the yield include vector's l -th component is calculated as $X F l = \max_{k, l} h_{k, l}$ $n k = 1$ k . [4]

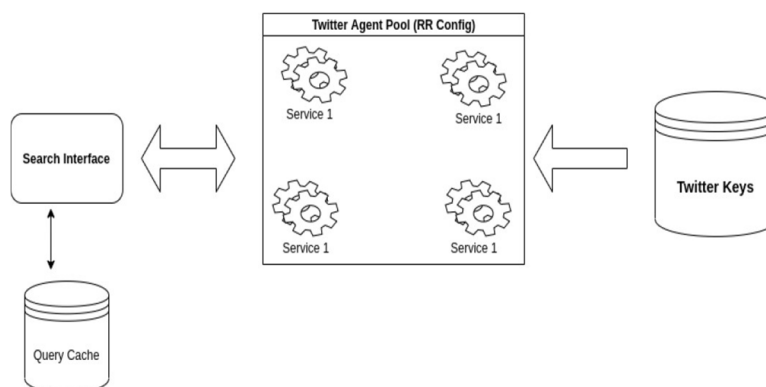


Fig.2.1:Twitter Service Design

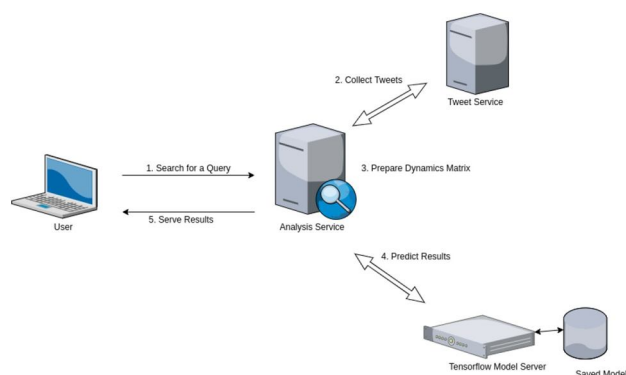


Fig.2.2: Integrated Service

- b) *A Logistic Regression:* In order to predict the probability of an absolute dependent variable, a Machine Learning method is used. The dependent variable is a parallel variable in this method that encodes information as 1 (accepted) or 0. (rejected) $P(Y=1)$ is predicted to be a part of X by the logistic regression model. Logistic regression is a statistical model in which the response variable has a discrete value and the descriptive variables can be continuous or discrete. In this case, the results may be true news ($Y = 1$) or fake news ($Y = 0$). The binary logistic regression model gives the probability that a record belongs to a positive class, $P(Y = 1)$, as follows:

$$P(Y=1) = e^z / (1 + (e^z))$$

- c) *Sigmoid Function:* The logistic function, also known as the sigmoid function, was developed by analysts to indicate population growth characteristics in the environment and it increases rapidly and maximally at the transport boundary in the environment. Any real esteemed number may be steered into a value between 0 and 1, but never exactly at those cutoff points, according to an S-shaped bend.

$$y = e^{(b_0 + b_1 * x)} / (1 + e^{(b_0 + b_1 * x)})$$

Here, y denotes the expected output, b_0 denotes the slope or capture term, and b_1 denotes the coefficient for an input value (x). An associated b coefficient (consistent actual value) should be discovered by using the training data for each section of your information.

Essentially, we fit the model to the preparation set, anticipate test set outcomes and calculate accuracy, precision and recall just as tune hyperparameters for ideal outcomes and accuracy.[5]

The proposed framework scratches numerous news sites and checks for keywords referenced in the info news story and computes a rate joining the outcomes from the AI approach which utilizes the calculated relapse calculation and the machine learning approach. Additionally, a central point to say something in terms of phony news is the feeling or the feeling behind the news story, the news story might be one-sided concerning particular political promulgation, thus for aside[5]. Yet, genuine news should be fair; it should be authentic, not blaming or definitive. Subsequently, to educate the client about the inclination in the news story, the proposed framework performs wistful examination including pre-preparing, NLTK lemmatization, watchman stemmer, and utilizations guileless Bayesian calculation and afterward gives the client the feeling behind the news story.

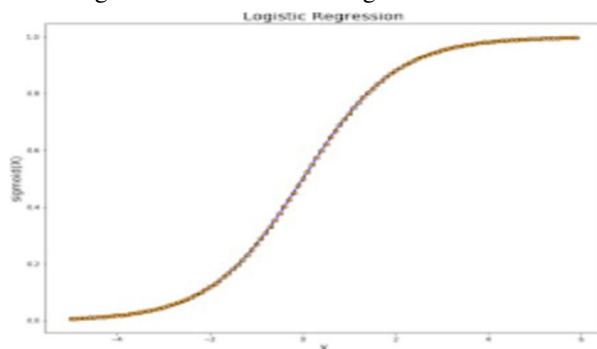


Fig.2.3: Logistic Regression

B. Random Forest Classifier

RF comprises an enormous number of decision trees working exclusively to predict a result of a class where the last expectation depends on whether a class that got a greater part casts a ballot. The mistake rate is low in random forests when contrasted with different models, because of low connection among trees. Our random forests model was prepared to utilize various boundaries; i.e., various quantities of assessors were utilized in a matrix search to create all that model that can foresee the result with high precision. There are many decision trees in a random forest, which is a grouping algorithm. In order to create an uncorrelated forest of trees whose advisory group's forecast is more accurate than any individual tree's, it uses bagging and emphasises arbitrariness while building each tree. There are many separate decision trees in a random forest, as the name implies. Using the Random forest, each tree emits a class expectation that we can use in our model. The class with the highest votes becomes the prediction. The random forest model's superior performance may be explained by the fact that it consists of many typically uncorrelated models (trees) working together as a board to outperform any of the individual component models. As a result, how can random forest ensure that the behaviour of any tree in the model isn't too linked to the behaviour of any other trees? Both the above methods are employed: [6] [7]

First, a bag is needed (Bootstrap Aggregation) Decision trees are very sensitive to the data they are provided.

Even little adjustments to the toolkit may result in quite complex tree topologies in the end. Random forest takes use of this by allowing each tree to erroneously test from the dataset using substitution, resulting in a wide variety of trees. This interaction is known as stowing or bootstrapping.[8]

1) Feature Randomness In an ordinary decision tree, when the time has come to part a hub, Then we look at all the possibilities and choose the one that provides the greatest separation between the perceptions in the left node and those in the right hub. Each random tree may only choose from an irregular subset of high points, in contrast to the predetermined forest structure. This increases the diversity of the trees in the model and, as a result, reduces the connection between them, leading to greater improvement. [9] Our study explores diverse textual properties that could be utilized to separate fake news from genuine. By utilizing those properties, we train a mix of various AI calculations utilizing different gathering techniques that are not completely investigated in the current literature. Many applications show that ensemble learners are valuable, since learning models like bagging and boosting tend to reduce mistake rates and increase accuracy. These procedures work with the preparation of various AI calculations in a viable and productive way. We likewise directed broad tests on 4 genuine world openly accessible datasets. The outcomes approve the improved presentation of our proposed strategy utilizing the 4 normally utilized exhibition measurements (in particular, accuracy, precision, recall, and F-1 score).[10]

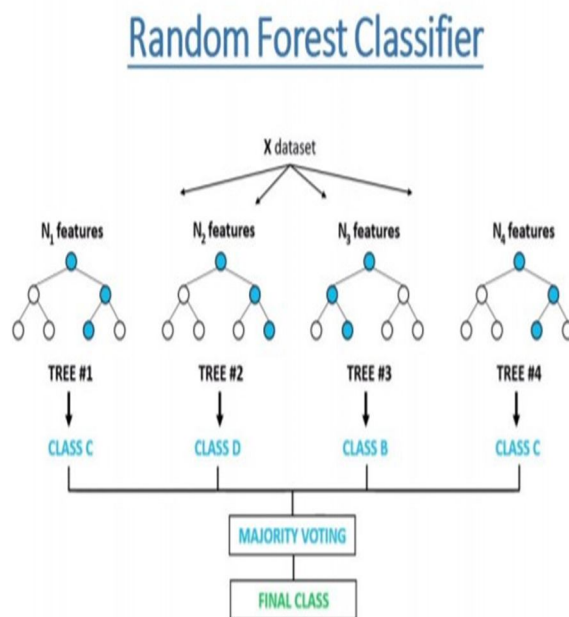


Fig.2.4:Random Forest Classifier

C. Decision Tree Classifier

A decision tree is a bunch of decision nodes beginning at the root. The advantages of using a decision tree incorporate simple understanding, effective treatment of anomalies, no requirement for the straight partition of classes, and subordinate highlights. In any case, the presence of so numerous meager highlights could lead a decision tree to overfit, and subsequently, it performs inadequately.

The J48 computation is perhaps the most well-known of the calculations used in the grouping.. It's all based on the C4.5 computation, which stipulates that all input must be quantitative and plain. A permanent type of data will thus be omitted from examinations. There are two distinct pruning techniques used by J48. Using a method known as subtree substitution, the main goal is to reduce the number of tests required by substituting hubs within a decision tree. The elevation of the subtree has a modest impact on the choice tree models in general. Most of the time, there is no exact way to predict the utility of a choice, even though it is wise to switch it off if the acceptance technique takes longer due to the subtree's rising being somewhat computationally confused.

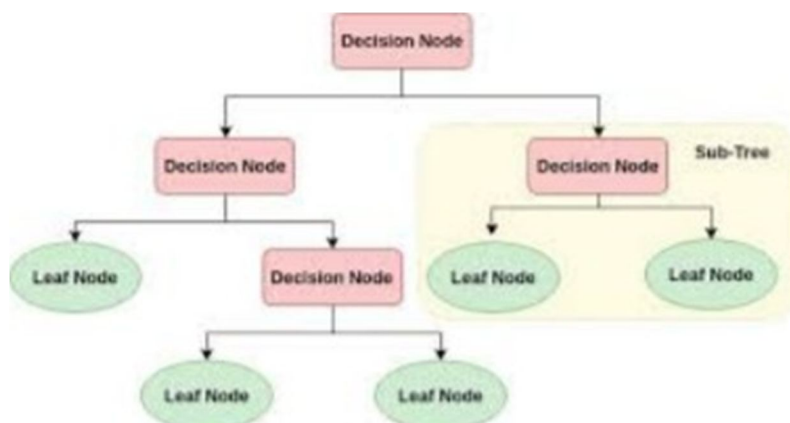


Fig.2.5.Decision Tree Classifier

D. Gradient Boosting Classifier

In machine learning, a technique known as gradient boosting may be used to solve regression issues by creating an expectation. Gradient boosting trees are used when a decision tree proves to be a poor learner. These trees usually outperform random forests. A self-assertive differentiable loss function may be streamlined using this method, which builds the model in the same phase-insightful manner as previous boosting methods.

Gradient Boosting is a training method that slowly and gradually improves the performance of many models. Gradient boosting model and Adaboost both select the weaknesses of weak learners, which is an important similarity (E.g. Choice trees). Even while the AdaBoost model detects the absence of high-weight information primarily on utilising gradients inside the loss function, AdaBoost boosts the performance by employing gradient information. The loss function serves as an estimate of how well the model's coefficients match the hidden information or data, showing how smart the model is. [11] [12]

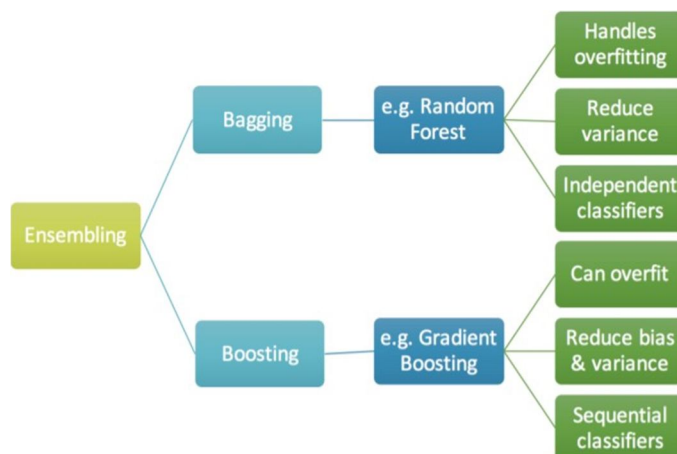


Fig.2.6:Gradient Boosting Classifier

III. EXPERIMENTAL DATA AND RESULTS

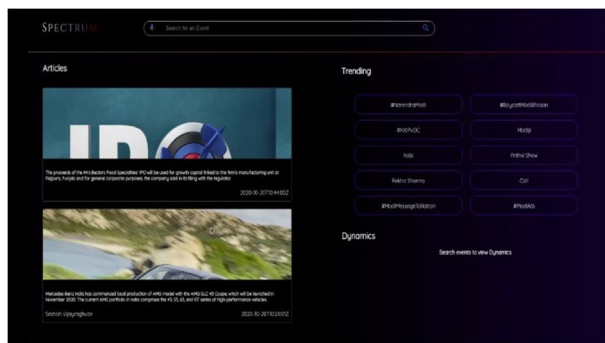


Fig 3.1 Home page

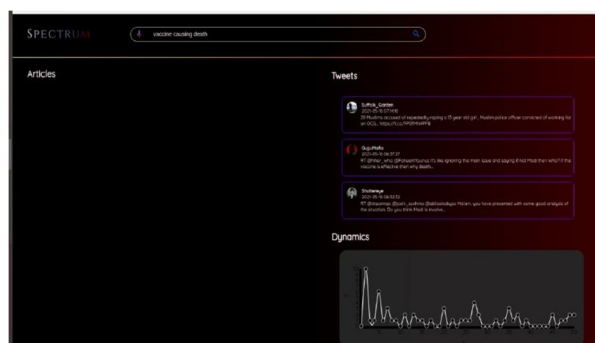


Fig 3.2 Fake News (denoted by a red shade on the X-axis-No. of hours, Y-axis-No. of tweets.)
X-axis-No. of hours, Y-axis-No. of tweets.

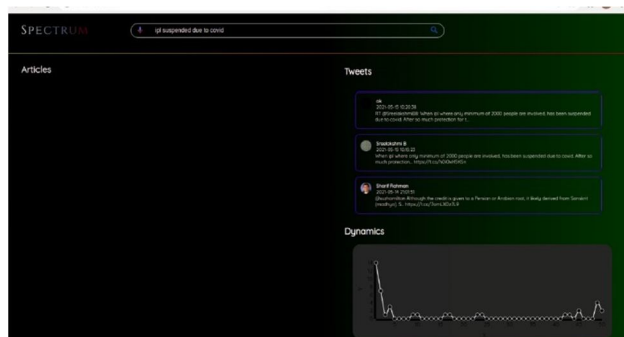


Fig 3.3 Real news (denoted by green shade on the right)
X-axis-No. of hours, Y-axis-No. of tweets.

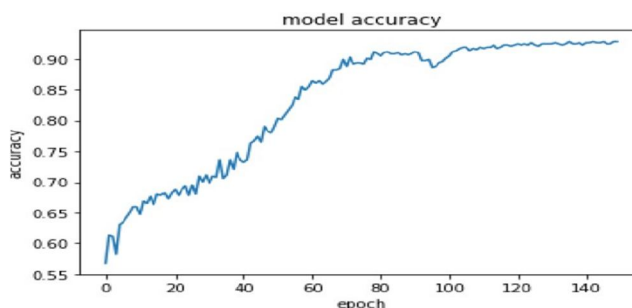
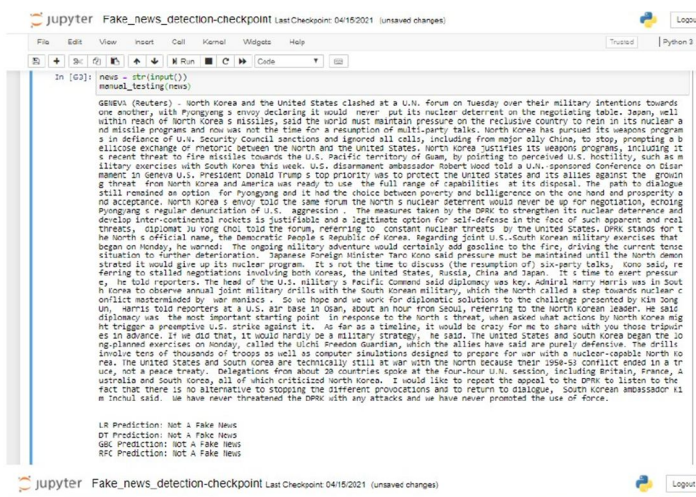


Fig 3.4 Epoch--Unit time
As time goes by, the accuracy of this model increases.

IV. ANALYSIS OF THE ALGORITHMS (LOGISTIC REGRESSION, DECISION TREE CLASSIFICATION, GRADIENT BOOSTING CLASSIFIER, RANDOM FOREST CLASSIFIER)

A. (Dataset from Kaggle)



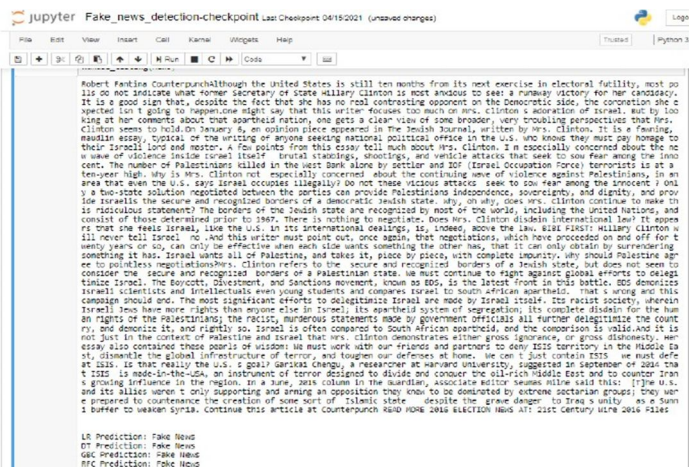
```

In [63]: news = str(input())
          manual_testing(news)

GENEVA (Reuters) - North Korea and the United States clashed at a U.N. forum on Tuesday over their military intentions towards one another, with Pyongyang's envoy declaring it would never put its nuclear deterrent on the negotiating table. Japan, well within reach of North Korea's missiles, said the world must maintain pressure on the nuclear country to rein in its nuclear and missile programs and now was not the time for a resumption of multi-party talks. North Korea has pursued its weapons program in defiance of U.N. Security Council sanctions and ignored the world's nuclear deterrent would never be up for negotiation, according to Pyongyang's regular denunciation of U.S. aggression. The measures taken by the DPRK to strengthen its nuclear deterrence and develop inter-continental rockets is justifiable and a legitimate option for self-defense in the face of such apparent and real threats, diplomat Ju Yong Chol told the forum, referring to constant nuclear threats by the United States. DPRK stands for the North's official name, the Democratic People's Republic of Korea. Regarding joint U.S.-South Korean military exercises that began on Monday, he warned: The ongoing military adventure would certainly, and leading to the fire, driving the current tense situation to further deterioration. Japanese Foreign Minister Fumio Kishida said pressure must be maintained until the North stated it would give up its nuclear program. It's not the time to discuss (the resumption of) six-party talks, Kishida said, referring to stalled negotiations involving both Korea, the United States, Russia, China and Japan. It's time to exert pressure, he told reporters. The head of the U.S. military's Pacific Command said diplomacy was key. Admiral Harry Harris was in South Korea to observe annual joint military drills with the South Korean military, which the North called a step towards nuclear conflict masterminded by war maniacs. We hope and we work for diplomatic solutions to the challenges presented by Kim Jong Un, Harris told reporters at a U.S. Air Base in Guam, about an hour from Seoul, referring to the North Korean leader. He said diplomacy was the most important starting point. In response to the North's threat, when asked what actions by North Korea might trigger a preemptive U.S. strike against it. As far as a timeline, it would be crazy for me to share with you those triggers in advance. If he did that, it would hardly be a military strategy, he said. The United States and South Korea began the 2019 joint military exercises on Monday, called the Ulchi Freedom Guardian, which the allies have said are purely defensive. The drills involve tens of thousands of troops as well as computer simulations designed to prepare for war with a nuclear-capable North Korea. The United States and South Korea are tactically still at war with the North because their 1953 ceasefire ended in a truce, not a peace treaty. Delegations from about 20 countries spoke at the four-hour U.N. session, including Britain, France, Australia and South Korea, all of which criticized North Korea. I would like to repeat the appeal to the DPRK to listen to the fact that there is no alternative to stopping the different provocations and to return to dialogue, South Korean ambassador Kim Chol said. We have never threatened the DPRK with any attacks and we have never prompted the use of force.

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 3.5: News article is Real.
(all 4 algorithms declared the news as real)



```

Robert F. Kennedy CounterpartAlthough the United States is still ten months from its next exercise in electoral facility, most people do not indicate any more serious fear of state military action is most anxious to see a runaway victory for him (candidate). It is a good sign that, despite the fact that she has no real contrasting opponent on the Democratic side, the competition she expected to be going to happen. Right up to this writer focused on much more Mrs. Clinton's approval of Israel, but looking at her comments about that apartheid nation, one gets a clear view of some broader, very troubling perspectives that Mrs. Clinton seems to hold. On January 8, an opinion piece appeared in the Jewish Journal, written by Mrs. Clinton. It is a far-right, misandrist essay, typical of the writing of anyone seeking national political office in the U.S. who knows they must pay homage to their Israeli love and master. A few points from this essay tell much about Mrs. Clinton. It is especially concerned about the new wave of violence in the Middle East. Israeli terrorism, she wrote, is a real threat. It is a real threat to the lives of the innocent. The number of Palestinians killed in the West Bank alone by settlers and IDF (Israel Occupation Force) terrorists is at a ten-year high. Why is Mrs. Clinton not especially concerned about the continuing wave of violence against Palestinians, in an area that even the U.S. says Israel occupies illegally? Do not these vicious attacks seem to you fear among the innocent? I only a two-state solution negotiated between the parties can provide Palestinians independence, sovereignty, and dignity, and provide Israel with the secure and recognized borders of a democratic Jewish state. Why, in only, does Mrs. Clinton continue to make this ridiculous statement? The borders of the Jewish state are recognized by most of the world, including the United Nations, and consist of those determined prior to 1967. There is nothing to negotiate. Does Mrs. Clinton disown international law? It is not that the Jews are Israel, like the U.S. in its international dealings, is, indeed, above the law. What else? Military Clinton will never tell Israel, no. And this writer must point out, once again, that negotiations, which have proceeded on and off for twenty years or so, can only be effective when each side wants something other than, that it can only obtain by surrendering something it has. Israel wants all of Palestine, and takes it, piece by piece, with complete impunity. Why should Palestine agree to Palestinian negotiations, Clinton refers to the secure and recognized borders of a Jewish state, but does not seem to consider the secure and recognized borders of a Palestinian state. We must continue to fight against global efforts to delegitimize Israel. The Boycott, Divestment, and Sanctions movement, known as BDS, is the latest front in this battle. BDS demands Israeli scientists and intellectuals even young students and compares Israel to South African apartheid. That's wrong and this campaign should end. The most significant efforts to delegitimize Israel are made by Israeli itself. Its racist society, wherein Israeli Jews have more rights than anyone else in Israel, its apartheid system of segregation, its complete disdain for the human rights of the Palestinian, the racist, murderous statements made by government officials all further delegitimizing the country, and despite it, and rightly so, Israel is often compared to South African apartheid, and the comparison is valid and it is not just in the context of Palestine and Israel that Mrs. Clinton demonstrates either gross ignorance, or gross dishonesty. Her essay also contains these pearls of wisdom: We must work with our friends and partners to deny Israel territory in the Middle East, dismantle the global infrastructure of terror, and toughen our defenses at home. We can't just contain ISIS. We must defeat ISIS. Is that really the U.S.'s goal? Caricatured Obama, who was a warmonger, suggested in September of 2011 that the U.S. is made-in-the-USA, an instrument of terror designed to divide and conquer the oil-rich Middle East and to counter Iran's growing influence in the region. In June, 2011, Obama, associate editor said this: I type the U.S. and its allies weren't only supporting and arming an opposition they knew to be dominated by extreme sectarian groups; they were prepared to counterbalance the creation of some sort of Islamic state despite the grave danger to Iraq's unity. As a Suni I suffer to weaken Syria. Continue this article at Counterpunch. How does this article read at 21st century with these lies?

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

Fig 3.6 News article is Fake
(all 4 algorithms declared the news as fake)

```

1. Logistic Regression

In [30]: from sklearn.linear_model import LogisticRegression

In [31]: LR = LogisticRegression()
          LR.fit(Xv_train,y_train)

C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
FutureWarning]

Out[31]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                             intercept_scaling=1, l1_ratio=None, max_iter=100,
                             multi_class='warn', n_jobs=None, penalty='l2',
                             random_state=None, solver='warn', tol=0.0001, verbose=0,
                             warm_start=False)

In [32]: pred_lr=LR.predict(Xv_test)

In [33]: LR.score(Xv_test,y_test)

Out[33]: 0.99528770053476

In [34]: print(classification_report(y_test, pred_lr))

              precision    recall  f1-score   support

     0       0.99       0.98       0.99       5919
     1       0.98       0.99       0.99       5301

 accuracy          0.99       0.99       0.99       11220
 macro avg          0.99       0.99       0.99       11220
 weighted avg          0.99       0.99       0.99       11220
  
```

Fig 3.7 Logistic Regression accuracy is 0.99

```

2. Decision Tree Classification

In [35]: from sklearn.tree import DecisionTreeClassifier

In [36]: DT = DecisionTreeClassifier()
          DT.fit(Xv_train,y_train)

Out[36]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                                max_features=None, max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, presort=False,
                                random_state=None, splitter='best')

In [37]: pred_dt = DT.predict(Xv_test)

In [38]: DT.score(Xv_test,y_test)

Out[38]: 0.996167557932639

In [39]: print(classification_report(y_test, pred_dt))

              precision    recall  f1-score   support

     0       1.00       1.00       1.00       5919
     1       1.00       1.00       1.00       5301

 accuracy          1.00       1.00       1.00       11220
 macro avg          1.00       1.00       1.00       11220
 weighted avg          1.00       1.00       1.00       11220
  
```

Fig 3.8 Decision Tree Classifier accuracy is 1.00


```

3. Gradient Boosting Classifier
In [46]: from sklearn.ensemble import GradientBoostingClassifier

In [47]: gbc = GradientBoostingClassifier(random_state=0)
gbc.fit(xv_train, y_train)

Out[47]: GradientBoostingClassifier(criterion='friedman_mse', init=None,
learning_rate=0.1, loss='deviance', max_depth=3,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=100,
n_iter_no_change=None, presort='auto',
random_state=0, subsample=1.0, tol=0.0001,
validation_fraction=0.1, verbose=0,
warm_start=False)

In [48]: pred_gbc = gbc.predict(xv_test)

In [49]: gbc.score(xv_test, y_test)

Out[49]: 0.9952762923351258

In [50]: print(classification_report(y_test, pred_gbc))

              precision    recall  f1-score   support

    0         1.00      0.99      1.00       5919
    1         0.99      1.00      1.00       5301

 accuracy          1.00      1.00      1.00      11220
 macro avg          1.00      1.00      1.00      11220
 weighted avg          1.00      1.00      1.00      11220

```

Fig 3.9: Gradient Boosting Classifier accuracy is 1.00

```

4. Random Forest Classifier
In [57]: from sklearn.ensemble import RandomForestClassifier

In [58]: RFC = RandomForestClassifier(random_state=0)
RFC.fit(xv_train, y_train)

Out[58]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
max_depth=None, max_features='auto', max_leaf_nodes=None,
min_impurity_decrease=0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, n_estimators=100,
n_jobs=None, oob_score=False, random_state=0, verbose=0,
warm_start=False)

In [59]: pred_rfc = RFC.predict(xv_test)

In [60]: RFC.score(xv_test, y_test)

Out[60]: 0.9769062218338681

In [61]: print(classification_report(y_test, pred_rfc))

              precision    recall  f1-score   support

    0         0.97      0.99      0.98       5919
    1         0.99      0.96      0.98       5301

 accuracy          0.98      0.98      0.98      11220
 macro avg          0.98      0.98      0.98      11220
 weighted avg          0.98      0.98      0.98      11220

```

Fig 3.10 Random Forest Classifier accuracy is 0.98

V. COMPARING THESE ALGORITHMS

	Logistic Regression	Decision Trees	Gradient Boosting Classifier	Random Forest Classifier	Propagation Dynamics
Precision	0.99	1.00	1.00	0.98	0.96
F1-Score	0.99	1.00	0.99	0.96	N/A
Recall	0.99	0.99	1.00	0.94	N/A
Avg_Train_Time	8 sec	40 sec	331 sec	12 sec	56 sec
Epoch	N/A	N/A	N/A	N/A	15

TABLE I.

A. Results in Brief

The four algorithms named Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and Gradient Boosting Classifier check the data input from a huge dataset which we retrieved from Kaggle roughly consisting of around 40000 (till the year 2017) with the equal number of fake and real news. At each step, for each algorithm, we try to define the accuracy of each algorithm for the outcome we need. We not only try to define the accuracy of the algorithms but also recall the F1-score which collectively defines the performance of that algorithm.

On the other hand, in Propagation Dynamics, we have made a webpage that extracts the tweets in real-time based on the query. We have used Docker and docker-compose which helps us to package applications into containers. We have implemented this in ubuntu. This model comes up with an output where we check the user input of news with various tweets regarding that topic and confirm whether the news input is true or not. And also, it shows us what are tweets it went through to verify whether the news is true or fake. It shows us a graph as well in which it tells us how many tweets have been posted per hour about the topic which we have searched about.

VI. HARDWARE SPECIFICATION

- 1) Intel i7-7100U CPU 10th Gen Processor
- 2) Clock Speed @2.40GHz
- 3) RAM 8.00 GB
- 4) Intel HD Graphics 620

A. Software Requirements

- 1) Microsoft Windows 10 Education 64-bit
- 2) Jupyter(Anaconda)
- 3) Docker and Docker Compos
- 4) Ubuntu 18.10

VII. CONCLUSION

In this project, we implemented 5 algorithms in total for detecting fake news-Logistic Regression, Decision Tree Classification, Random Forest Classifier, Gradient Boosting Classifier, and Propagation Dynamics.

Looking into the various algorithms we implemented, we found that the first four i.e., Logistic Regression, Decision tree classification, Gradient Boosting Classifier, and Random forest classifier always give a definite value(i.e-Fake or real)without further justification for the same. It is extremely challenging in showing the justification for these three, and even if it is done, it would give a very less accuracy of the result. Keeping it aside, these 4 algorithms come up with great accuracy in their result(almost 99%) but the gradient Boosting Algorithm takes a huge amount of time as compared to others in providing the result.

Whereas the fifth algorithm we used i.e., Propagation Dynamics not only gives the definite result but also can give proper justification for the same(Here, in this case, we are using tweets to verify whether the news is true or not). It also shows the propagation of the news we entered in the search box. The only disadvantage to this algorithm is that it comes up with not-so-good accuracy(96% accuracy).

We have successfully built a machine learning model that can predict the authenticity of news articles or events with 96% accuracy. The model collects tweets in real-time and analyses the propagation dynamics of the rumored events to determine their authenticity. The model has been served publicly as a REST API for better reach. To include local rumors and events in analysis, the system needs to broaden the medium of collection to local newspapers and news websites. The system also needs to include a feedback mechanism to include people's views when deciding on an event.

REFERENCES

- [1] Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumors in social media: A survey," ACM Computing Surveys, vol. 51, pp. 32, 2018.
- [2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explorations Newsletter, vol. 19, pp. 22–36, 2017.
- [3] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," PloS one, vol. 10, 2015.
- [4] V.L. Rubin and T. Lukoianova, "Truth and deception at the rhetorical structure level," JASIST, vol. 66, pp. 905–917, 2015.
- [5] D. M. Nguyen, T. H. Do, R. Calderbank, and N. Deligiannis, "Fake news detection using deep Markov random fields," in NAACL, 2019, pp. 1–10.
- [6] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in AAAI, 2016, pp. 2972–2978.
- [7] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on Sina Weibo by propagation structures," in ICDE, 2015, pp. 651–662.
- [8] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in WWW, 2011, pp. 675–684.
- [9] J. Ma, W. Gao, P. Mitra, S. Kwon, B. Jansen, K. F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in IJCAI, 2016, pp. 3818–3824.
- [10] J. Ma, W. Gao, Z. Wei, Y. Lu, and K. F. Wong, "Detect rumors using time series of social context information on microblogging websites," in CIKM, 2015, pp. 1751–1754.
- [11] A. Magdy and N. Wanas, "Web-based statistical fact-checking of textual documents," in Int. workshop on Search and mining user-generated content, 2010, pp. 103–110.
- [12] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "Fake review detection: Classification and analysis of real and pseudo reviews," UIC- CS-03-2013. Technical Report, 2013.
- [13] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: On political fact-checking and fake news," in EMNLP, 2017.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)