Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 22 Issue 2 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Fake News Detection: Covid-19 Perspective

By Md. Ziaur Rahman Shamim, Shaheena Sultana, Anika Tabassum, Israt Tabassum & Sarkar Binoyee Farha

Notre Dame University

Abstract- The development of social media has contributed to a remarkable rise in the spread of fake news. Today people rely more on online news outlets. The chance of receiving fake news on an online platform is high. As we went through a pandemic and the Covid-19 was the most absorbing topic of 2020, much news on Covid-19 was published every day in traditional media and social media. Among that news, some are fake. In this work, we have collected a new dataset for detecting fake news from traditional media on Covid-19. We have gathered more than 3000 pieces of news from traditional media out of the 170 are fake ones that were collected from fact-checking sites. Then we have tested the existing four classification algorithms with our dataset using Count Vectorizer and TF-IDF. We have merged 170 fake news with four scales of true news and analyzed the outcome.

Index Terms: fake news, fake news detection, traditional media, data set, covid-19, social media.

GJCST-C Classification: DDC Code: 004.678 LCC Code: TK5105.875.I57

Fake News Detection: Covid-19 Perspective

Md. Ziaur Rahman Shamim [α], Shaheena Sultana [σ], Anika Tabassum [ρ], Israt Tabassum [Ѡ]
& Sarkar Binoyee Farha[¥]

Abstract- The development of social media has contributed to a remarkable rise in the spread of fake news. Today people rely more on online news outlets. The chance of receiving fake news on an online platform is high. As we went through a pandemic and the Covid-19 was the most absorbing topic of 2020, much news on Covid-19 was published every day in traditional media and social media. Among that news, some are fake. In this work, we have collected a new dataset for detecting fake news from traditional media on Covid-19. We have gathered more than 3000 pieces of news from traditional media out of the 170 are fake ones that were collected from fact-checking sites. Then we have tested the existing four classification algorithms with our dataset using Count Vectorizer and TF-IDF. We have merged 170 fake news with four scales of true news and analyzed the outcome.

Index Terms: fake news, fake news detection, traditional media, data set, covid-19, social media.

I. Introduction

In 2020, we have gone through trauma due to covid-19. This pandemic is not less than a trauma. Covid-19 has become a challenge to each and everyone whose country is facing a huge number of positive cases regarding this and the number of deaths are rising on a daily basis, which actually makes one depressed. During this pandemic situation people are stuck in their own places thus they are consuming much more internet than before. According to openvault, data usage has increased 47 percent during this pandemic [34]. As people are consuming more internet, they are also getting their news from it. Among those news stories, some of them are fake, but it is hard for maximum people to find out whether the news is fake or not. Through an automated system we can easily detect fake news. An automated system could be defined as a technique that performs a task using programming inputs and computerized feedback control to verify that the instructions are followed correctly. The end result is a technology that can function without any need for human interaction [18]. An automated system needs a well-labeled data set to detect fake news from tons of news. So, we have made a new data set on Covid-19.

What is Fake News? Fake news is misleading or false news information introduced as news [19]. There is no valid definition of fake news. Many researchers [32], [11] have adopted Hunt Allcot et al. definition. They defined, a news article that is deliberately and verifiably false and could deceive readers is called fake news [4]. Also another explanation from S. Desai, fake news is those articles that are fabricated, false information with no supportable facts, origin, or quotes [13]. These two definitions have two key features: authenticity, fake news includes false information, facts, sources, or quotes to make it more authentic so that readers can have no doubt about the news and another is intent. Basically, fake news is generated with the purpose of deceiving readers, sometimes to harm individuals. In some papers, satire or parody news are also considered as fake news [7], [8]. Fake news can be spread via two types of media such as traditional media and social media [32]. Traditional media are newspapers, magazines, blogs, etc. and social media are Facebook, Twitter, WhatsApp, etc. So basically, we have focused our data set to detect fake news from traditional media and news related to Covid-19.

According to researchers, in the first three months of the year 2020 at least 800 people might have died across the world because of fake news about Covid-19 [10]. Fake news is not only spreading disinformation or rumors, but also taking valuable lives and also hampering public properties. To check the facts of fake news there are some delegated websites like Politifact [20], lead stories [15], FactCheck [28], the full fact [24], health feedback [35] where they manually updates potential fake news stories which are published online platforms.

We have collected a vast number of news data sets so that our outcome can help the upcoming researchers to work on this problem. We have worked on Covid-19 which is the most hooked and challenging topic in the world and every day we have got thousands of news regarding this pandemic. From those news some of them are fake. In the beginning, people used to believe in every news regarding Covid-19. By these, some of them became so panic that they were about to lose their lives and some of them already lost theirs. There were too many rumors that people started believing those. For example, in Iran, there were rumors across the country that alcohol consumption would help combat Covid-19 [16], [5].

Author α σ ρ Ѡ ¥: Department of Computer Science and Engineering, Notre Dame University Bangladesh Dhaka, Bangladesh.
e-mails: zrshamim8822@gmail.com, shaheenacse@ndub.edu.bd, anika4725@gmail.com, israt.tabassum34@gmail.com, farha1binoyee@gmail.com

Hence, people started consuming alcohol in large numbers. Like this news, rumors were being spread across the world and people started to believe that news. In this work, we have presented a new data set to detect fake news. For this data set, we have gathered more than three thousand news from various news sites. In our data set, there are two types of news, one is true news another one is fake news. There are one hundred and seventy fake news in our data set. For detecting fake news from our data set, we have used four classification algorithm they are multinomial Naïve bayes, logistic regeression, support vector machine and passive aggressive classifier.

The rest of the work is organized as follows: in chapter II, we have reviewed some of the previous work. In chapter III, we have discussed the algorithm that we have used for our work. In chapter IV, we have discussed our work. In chapter V, we have discussed the performance and have concluded our work in chapter VI.

II. Literature Review

Previously many researchers have worked on this particular area. Some of them created new models or explore new areas to detect fake news. Others improved the existing model to improve fake news detection and others build a robust data set to fuel the detection system. In this section, we explore some of the previous works. Ruchansky et al. presented a fake news detection System using a Hybrid Deep Model. They proposed a model called the CSI. Capture, Score, and Integrate are the three modules of CSI model. The capture module is based on text and reaction. This module captured the engagement between users and news articles, the second module score learns the source feature. Then they combined two modules and integrated them with the third module to produce a label for fake news [29]. Kai Shu et al. presented a brief analysis of detecting fake news from social media, they also discussed fake News Classification based on social context and psychology, also reviewed current algorithm in terms of data mining [32]. Lutzke et al. conducted a 3 by 2 experimental model having guidelines, enhanced guidelines, and a controlled part. Each experiment has two types of news about climate change one is fake and another is real. A total of 2750 people from different fields participated in their experiment. Each participant was randomly allocate to one of six possible experimental variations [22]. Bahad et al. used deep learning model, Bi-directional Long short-term memory (LSTM) over other techniques like as Convolutional Neural Networks (CNN), unidirectional LSTM, Recurrent neural network (RNN), for detecting fake news [6]. Abdullah et al. used machine learning and deep learning technique to detect fake news from Twitter data set. They have used five different Machine Learning algorithms, like Support Vector Machine, Naïve - Bayes algorithm, Logistic Regression, and Recurrent Neural Network to detect fake news from the data set [1]. Zhang et al. used a two-layer method that includes the identification of fake topics and fake incidents [38]. Yang et al. detect satirical news using linguistic features [37]. Apuke and Omar developed a complete model studying from uses and gratification view [14]. Agarwal et al. used the natural language processing and the machine learning method to detect fake news [2]. Granik and Mesyura used a Naïve- Bayes classifier to classify news from BuzzFeed data sets [17]. Yang presented Liar data set containing news more than 12 thousand for fake news detection. The data set is divided into six classes [36]. Fake News Corpus by Szpakowski. He uses multiple corpora to develop and test various models. The corpus has been automatically crawled using open sources.co labels [33]. Khan et al. performed benchmark research to compare the performance of several machine learning methods on three distinct data sets [21]. Zhang and Ghorbani gave a detailed review of what has been discovered so far on false news. They have described the detrimental impact of online fake news as well as the current status of detecting tools [39]. Alkhodair et al. presented a novel method for automatically identifying rumors, combining the learning of word embedding with the training of a recurrent neural network with two separate goals [3].

III. Algorithms

Previously many algorithms have been developed to detect fake news. We have used four different types of supervised learning algorithms to detect fake news. We briefly discuss the four existing detection algorithms in this section.

In machine learning, Supervised learning is the task of creating a function that maps an input to an output based on given input-output pairs [30]. In supervised learning, the program is given labeled input data and the output result is expected. In regression and classification problems, supervised learning works fine, such as deciding what group a news story belongs to.

a) *Multinomial Naïve Bayes*

Naïve bayes classifiers are members of "probabilistic classifiers" which is build from applied bayes theorem with strong (Naïve) independent assumptions between the attributes. Due to its simplicity, speed, and good accuracy Multinomial Naïve bayes [23] is a common classifier for text classification. Multinomial classifier learns a conditional probability that the nth document x^n form a class y_c given the document x^n that is formed by Bayes rule, $p(y_c|x^j) = \frac{p(y_c)p(x^j|y_c)}{p(x^j)}$, Multinomail Naïve bayes uses multinomial model to find outx^j $p(f_i|y_c)$. So, in perticular

$p(x^j|y_c) = \pi_{i=1}^{n_f} p(f_i|y_c)^{x_i^i} n_f$, is the number of feature x_i^i is the number of times the ith feature occurs in the jthdocumentx^j $p(f_i|y_c)$is the probability of the ith feature occurring given class y_c

In text classification, the conditional feature probabilities are estimated using Laplace smoothing. Laplace smoothing can be written as,

$$P_Laplace(f_i|y_c) = \frac{\lambda + \sum_{j=1}^{n_d} x_i^j}{\sum_{k=1}^{n_f}[\lambda + \sum_{j'}^{n_d} x_{j'}^k]}$$

Here, n_d is the number of data points in the training set from class y_c n_f , is the number of features λ, is a parameter known as the Laplace smoothing constant. λ is typically set to 1. As λ is a non-zero number, it prohibits such degenerate cases from equaling zero. In general, if a feature does not occur in any document of the training set, $p(x^j|y_c)$ or all documents x_j.The test set that contains the degenerated feature will be zero for all classes y_c, causing Multinomial Naïve bayes to lose all discriminating power.

b) *Logistic Regression*

Logistic regression is a classification model that models a binary variable using a logistic function [31]. It is a categorical dependent variable prediction approach that uses a collection of independent variables to forecast a dependent variable.

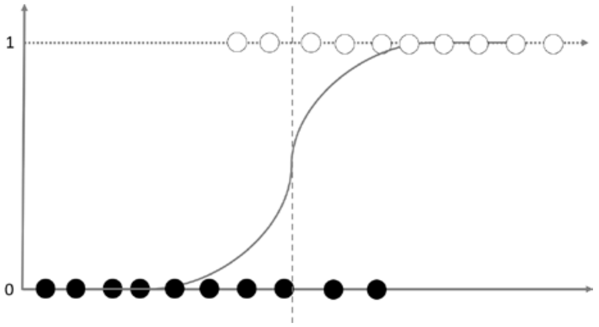


Fig.1: Example of a Logistic Regression

Because of the application of the sigmoid function in logistic regression, the curve in Fig. 1 is generated. The sigmoid curve is another name for the curve produced above.

c) *Support Vector Machine*

Vapnik and Cortes proposed a Support vector machine (SVM). The Support vector machine is a supervised machine learning method used for classification, regression problems. It determines the best decision boundary among various vectors [12]. Support Vector Machine draws a hyper-plane to separate two distinct classes [25].

Support Vector Machines not only create hyper-plane but also construct a maximum margin separator, a decision boundary, with the largest possible distance. SVM constructs the marginal line Fig. 2 by the nearest one or more positive or negative points, these points are called to be support vectors. The Marginal lines are drawn

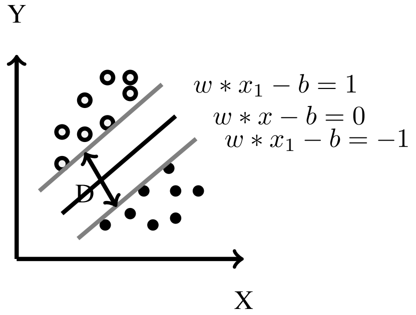
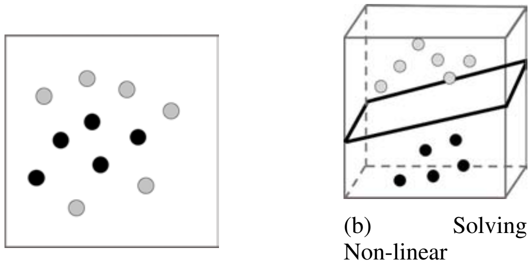


Fig. 2: Support Vector Machine

parallel with the hyper-plane and the distance between two marginal lines with the hyper-plane is equal. The total distance (D) between the two margins is called to be ma rginal distance. For having a better model SVM finds the largest distance (D) between support vectors in shorts marginal lines.

SVM produce a linear hyper-line of separation between classes which are linearly separable. But in nonlinear classification problems like as Fig. 3, it is not possible to have better classification only by drawing a hyper-plane. To solve Fig. 3a this kind of problem support vector machine using the so-called kernel trick.

They have the capability to implant the data from a lower dimension into higher-dimensional space. In Fig. 3b we can see that, the kernel trick takes the lower space input points Fig. 3a and then implant them into a higher dimension.



(a) Non-linear data point.

(b) Solving Non-linear classification problem using Kernel.

Fig. 3: A Non-linear classification problem.

d) *Passive Aggressive Classifier*

The passive-aggressive algorithm is a member of Machine learning algorithms. Passive-Aggressive algorithms are called so because :

- *Passive:* Keep the model and do not make any adjustments if the prediction is accurate.
- *Aggressive:* Make adjustments to the model if the prediction is incorrect.

The key idea of passive-aggressive classifier is that with each misclassified training data point it gets, the classifier changes its weight vector, attempting to fix it.

Initially, the weight W is 0, then the algorithm receives a new doc, then the algorithm applies feature extraction and normalizes the doc. If the dTw value is greater than 0 the algorithm will predict the class as a positive and prediction class will be $y = \pm 1$.

Algorithm 1 Passive Aggressive Classifier.

```
Initialize w = (0,....,0)
Receive a new doc $d = d_1.....d_v$
Apply feature extraction, normalize $\| d \| = 1$
Predict positive if $d^T w > 0$
Observe true class : $y = \pm 1$
Want to have:
$d^T w \geq +1$ if positive $(y = +1)$
$d^T w \leq -1$ if negative $(y = -1)$
Same as: y $d^T w \geq 1$
$Loss : L = max(0, 1 - y(d^T w))$
Update:$w_{new} = w + yLd$
```

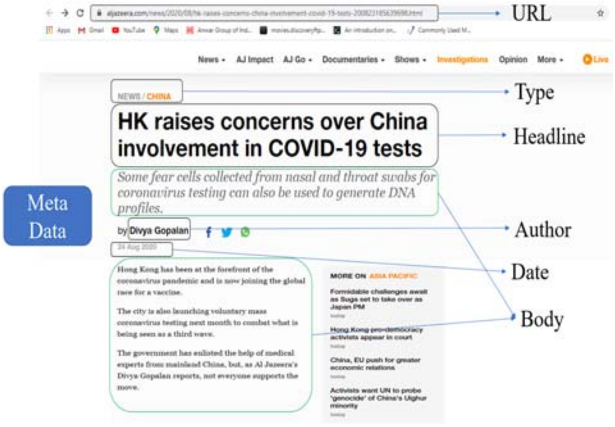
The core concept of the passive-aggressive classifier is to determine the loss function and act according to it. So, the loss function of the passive-aggressive classifier is Loss: L = max(0; 1 − y(dTw)) If there any misclassification, there will be a loss. If the data set has an input class that belongs to a positive class, but initially the output is a negative class. For this misclassification, the passive-aggressive classifier will update its initial weight.

IV. Our Work

A workflow is a sequence of tasks that processes a set of works. In Fig. 4 we have shown our workflow. Firstly, we have collected true and fake news from the various traditional media. As our data are text data so for feature extraction we approach through two different feature extractors one is TF-IDF (Term Frequency and Inverse Document Frequency) and another is count vectorizer. Then we have tested them with four different supervised algorithms and finally discussed the final result based on our data set.

a) Data Collection Process

The main focus of our work is to collect news from various traditional news media as shown in Fig. 6. We choose the most popular and reputable news websites from across the world to construct a collection of authentic news. Such as Aljazeera, The New York Times, BBC, CNN, and many other news websites as well. For fake news, we have collected from some reputed fact-checking websites such as Lead Stories, PolitiFact, Health Feedback, etc. These fact-checking websites offer a clear and instructive explanation of fake news that has previously been published on other websites. We have the nine metadata from news sites, the title of a news annotated as a headline, the main article referred to the body, URL of the news, the date of publication, publisher name, the type of the article. In the fake news data set, we added two additional attributes, one is a fact-checking website name and another one is URLs of fact articles for the authentication of the news. An example of the data collection process for our data set is shown in Fig. 5. Our collected data set is given in the link 1. We have labeled the true as 1 and the fake news as 0.



NEWS / CHINA

HK raises concerns over China involvement in COVID-19 tests

Some fear cells collected from nasal and throat swabs for coronavirus testing can also be used to generate DNA profiles.

by Divya Gopalan

Hong Kong has been at the forefront of the coronavirus pandemic and is now joining the global race for a vaccine.

The city is also launching voluntary mass coronavirus testing next month to combat what is being seen as a third wave.

The government has enlisted the help of medical experts from mainland China, but, as Al Jazeera's Divya Gopalan reports, not everyone supports the move.

MORE ON ASIA PACIFIC

Formidable challenges await as Suga set to take over as Japan PM

Hong Kong pro-democracy activists appear in court

China, EU push for greater economic relations

Activists want UN to probe 'genocide' of China's Uighur minority

Fig. 5: Meta data example

b) Data Preprocessing

Data preprocessing is a data mining approach that includes transforming raw data into a suitable state. Raw data is frequently incomplete, inaccurate, and even absent, and it is prone to have many faults in certain actions or patterns. Preprocessing data is a well-known approach for resolving such issues.

Initially, input data is typically provided in the raw state, which is in the texts, sentences, comments, articles, messages etc format. It requires certain clean-up or preprocessing before the data can be moved to machine learning algorithms, so algorithms can focus on key terms instead of terms that adds limited or no importance. So in raw state text data, there is a special character. Non-alphanumeric characters, as we all know, are special characters, these symbols are most frequently found in text data. These symbols contribute little meaning to text

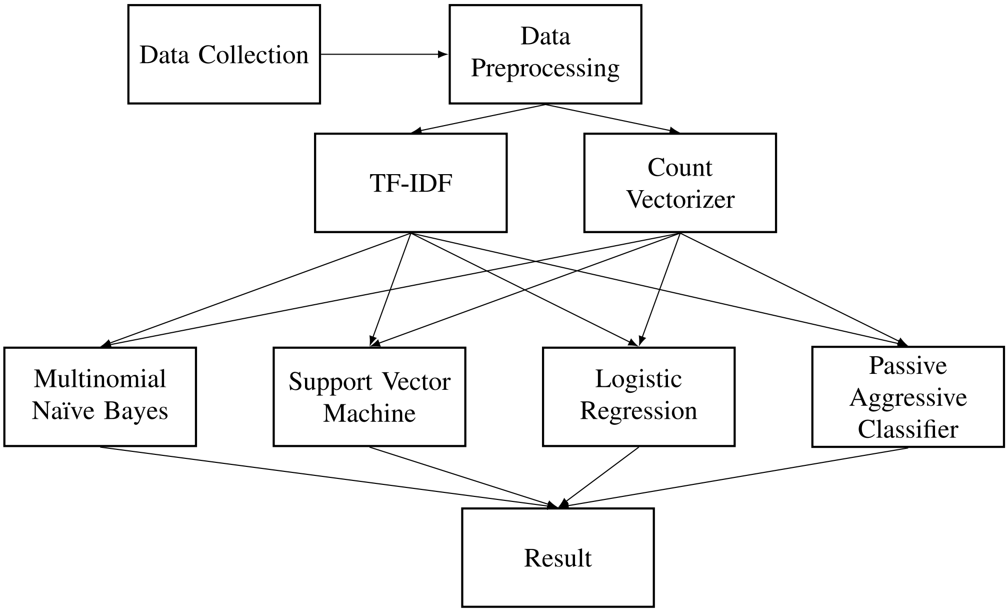


Fig. 4: Workflow of fake news detection: the Covid-19 perspective

noise. So, to avoid noise the whole raw data need to convert into a single form either all words into small character or capital character. For the further addition we have converted our data into a set of arrays to perform stemming and removing stop words.

c) Porter Stemmer

For our work, we have used porter stemmer for stemming. In 1980 a stemming method has gained much popularity which is now known as porter Stammer [26]. Speed and simplicity, are its identification. Data mining and Information retrieval are the main application of porter stammer. As porter stammer uses suffix stripping to produce stems. It produces the best output as compared to other stammers and also it has less error rate than other techniques. However, English words are limited to their application. The output of stem might not be a meaningful word, but the collection of stems is plotted onto the same stem as well.

d) Stop Words

There are certain words in a sentence as an example, be, too, not, etc. in the English language, which does not have any significance for the processing of natural language. So, during natural language processing, such words are taken out. In fact, stop words are some words that are stripped out of the processing of natural language [27]. In particular, stop words do not bring a lot of value to natural language processing results. Without modifying or compromising the context of every tatement, we can comfortably neglect it.

e) Feature Extraction

As our data is text data containing words. So, for use as inputs in machine learning algorithms, these terms now have to be represented as integers, or floatingpoint values. This method is considered the extraction of features or vectorization.

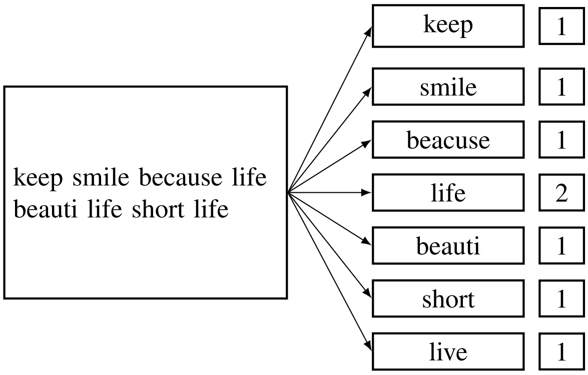


Fig. 8: Count Vectorizer

Count Vectorizer: The count vectorizer is often used to transform a text data collection to a word count vector. This transforms a set of textual data into a token count matrix. As we can see in Fig. 8 count vectorizer converts each word from text data into a count vector. Actually, the count vectorizer counts how frequently a

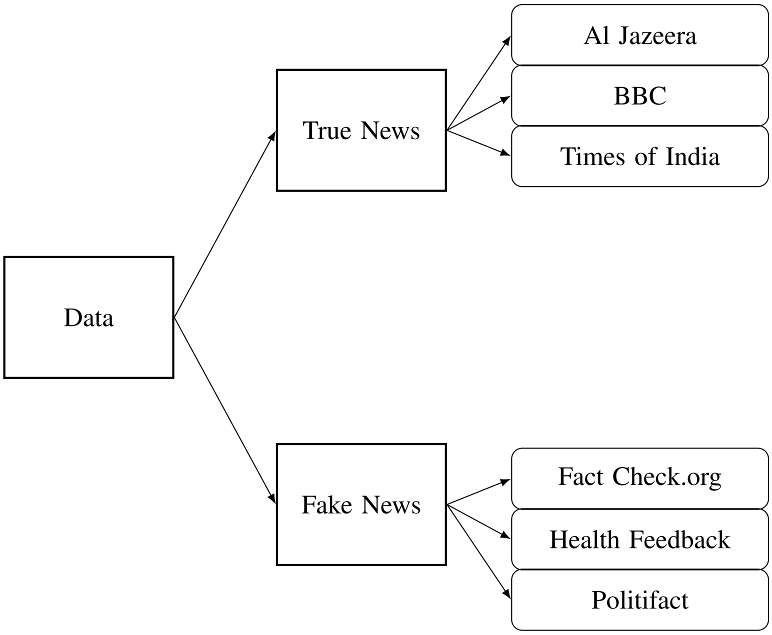


Fig. 6: Data collection process

- *Term Frequency(TF):* Term frequency calculates the frequency of a term or word occurs in a data set.

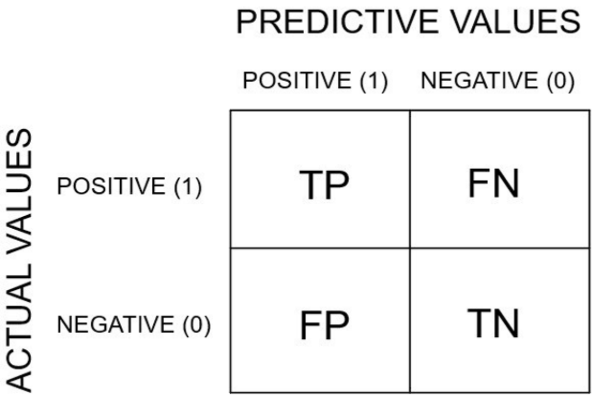
$$\text{TF(t)} = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}$$

- *Inverse document frequency(IDF):* Inverse document frequency calculates how important a word or term is in a document. In TF all words or terms are equally considered important. Thus, IDF weight down repeated words or terms and scale up the uncommon ones.

$$\text{IDF(t)} = log_e \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}$$

f) Confusion Matrix

A confusion matrix is a method of describing the classification algorithms performances [9]. A confusion matrix is a table that shows how well a classification model (or "classifier") performs on a testing data set for which the real values are known. The confusion matrix itself is uncomplicated, but the related terms might be complex.



		PREDICTIVE VALUES	
		POSITIVE (1)	NEGATIVE (0)
ACTUAL VALUES	POSITIVE (1)	TP	FN
	NEGATIVE (0)	FP	TN

Fig. 9: Confusion Matrix

- *True Positive(TP):* When the positive type is predicted exactly by a model.
- *True Negative(TN):* When the negative type is predicted exactly by a model.
- *False Positive(FP):* When the positive type is predicted incorrectly by a model.
- *False Negative(FN):* When the negative type is predicted incorrectly by a model.
- *Accuracy:* Accuracy is the ratio of accurate forecasts to overall forecasts made.

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}}$$

- *Precision:* Precision is the ratio of positive results

- *Precision:* Precision is the ratio of positive results correctly predicted to all positive results predicted.

$$\text{Precison} = \frac{\text{TP}}{\text{TP + FP}}$$

- *Recall:* Recall is the ratio of rightly predicted positive observation of all outcomes in actual class-yes.

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}}$$

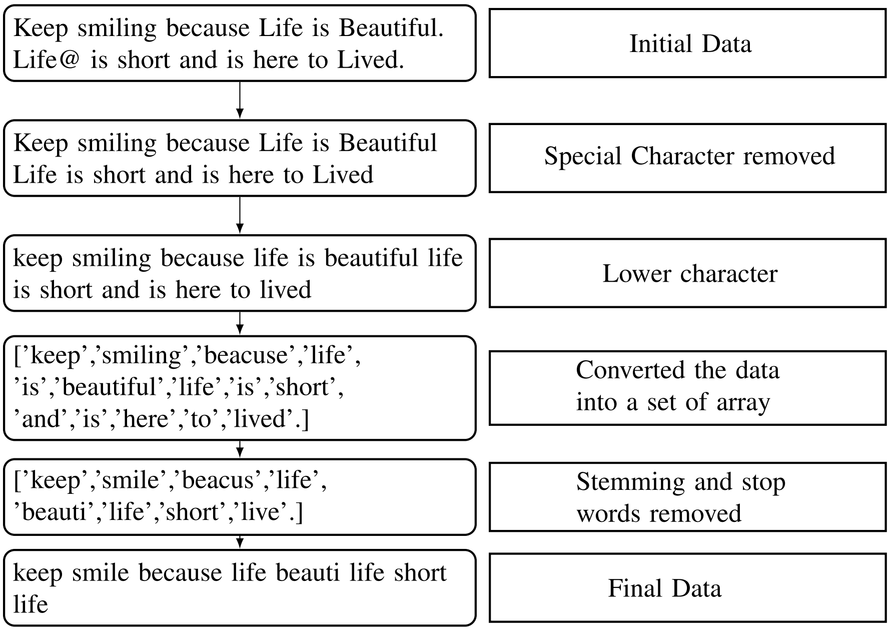

Keep smiling because Life is Beautiful. Life@ is short and is here to Lived.
Initial Data
Keep smiling because Life is Beautiful Life is short and is here to Lived
Special Character removed
keep smiling because life is beautiful life is short and is here to lived
Lower character
['keep','smiling','beacuse','life', 'is','beautiful','life','is','short', 'and','is','here','to','lived'.]
Converted the data into a set of array
['keep','smile','beacus','life', 'beauti','life','short','live'.]
Stemming and stop words removed
keep smile because life beauti life short life
Final Data

Fig. 7: Data Preprocessing

- *F1-Score:* F1-Score is the weighted average of Precision and Recall.

$$\text{F1-Score} = \frac{2 * \text{Pre} * \text{Re}}{\text{Pre + Re}}$$

V. Performance Evolution

We tested the detection system using two different types of feature extraction one is a count vectorizer and another is TF-IDF. As we have only one hundred seventy fake news so we merged them with four different scales (Table 1) of true data to have a depth look at the detection system.

Table 1: Mixing Fake Data with Four Different Scales of True Data.

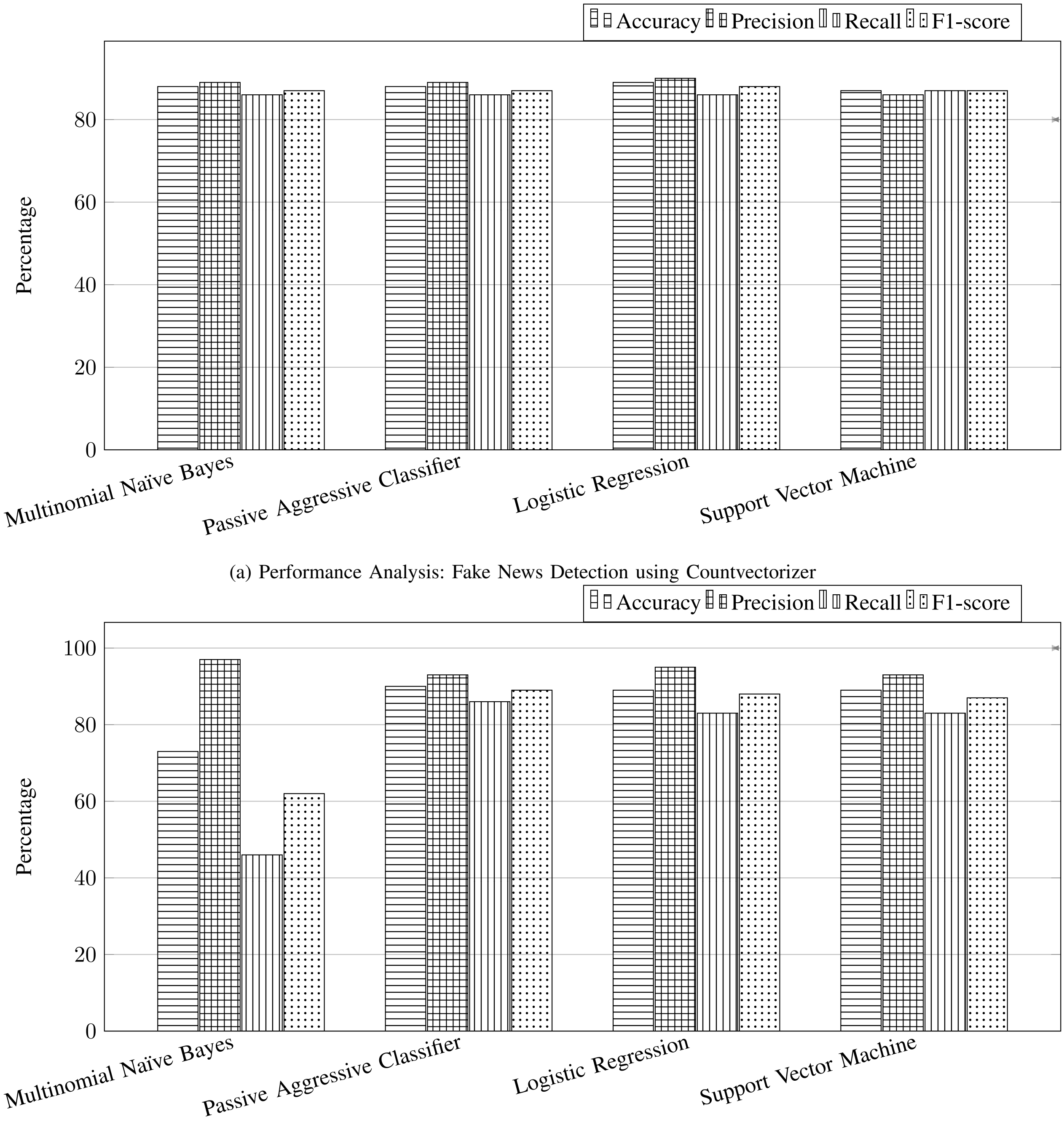
	Fake News	True News	Total
Set-1	170	180	350
Set-2	170	230	400
Set-3	170	330	500
Set-4	170	530	800

In set-1 we have merged 170 fake news with 180 true news. In set-2 we have 80 more true news than fake news, 58% of total data is true news. In set-3, 66% of data is truly labeled, and set-4 having 76% of true news. Then we have divided our data set into the training set and test set.

First, we have Multinomial Naïve bayes using TFIDF. We can see our final result in Table 2. Multinomial Naïve bayes shows the highest test accuracy of 89% on set-1 than other three algorithms using TF-IDF. Also shows stable values on precision, recall, and f1-score in fake news detection. In other sets, the test accuracy of Multinomial Naïve bayes declined as the amount of true data is much greater than the fake data. In set 4 it seems that the accuracy of 81% is much better than the previous two sets. But we can see that the recall (8%) and f1-score (14%), which is too much low in detecting fake news. In Table 3, we have our final result of Multinomial Naïve bayes using count vectorizer. We have a different scenario than TF-IDF while using a count vectorizer. In Table 3, we can see that the recall and f1-score drastically decreased while using TF-IDF. But in the count vectorizer, the accuracy increased as expected but the other three values did not drastically decrease shown stable value 3. In set 2 and set 3 Multinomial Naïve bayes shown the same accuracy of 88% (Table 3).

On set 4 Passive aggressive classifier shows highest test accuracy of 92% using TF-IDF than other three sets. But the poor value in recall (64%) and f1-score (76%). But in set 2 and set 3 it gives better results on all sections. Passive aggressive classifier using count vectorizer on Table 3. In this part set 2 and set 4 both shown the highest accuracy of 90% then other two sets. But set 4 has a poor outcome on recall (68%) and f1-score (73%) while detecting fake news than other sets, that we have tested. The other three set shown decent outcomes on all three parameters.

Logistic regression also given the highest accuracy of 91% in set 4. But not a satisfying outcome on the other three-parameter in detecting fake news. In general, set 2 given the best performance out of other four sets, having an accuracy of 89%. In set-4, Fake news detection



(a) Performance Analysis: Fake News Detection using Countvectorizer

(b) Performance Analysis: Fake News Detection using TF-IDF

Fig. 10: Fake News Detection Comparison Between Four Algorithm Using Set-2

Table 2: Fake News Detection Result Using TF-IDF

Technique	Algorithms		T/F	Support	Precision	Recall	F1-Score	Accuracy
TF-IDF	Multinomial naïve bias	Set-1	0	53	0.93	0.81	0.87	0.89
			1	63	0.86	0.95	0.9	
		Set-2	0	63	0.97	0.46	0.62	0.73
			1	69	0.67	0.99	0.8	
		Set-3	0	61	1	0.23	0.37	0.72
			1	104	0.69	1	0.82	
		Set-4	0	53	1	0.08	0.14	0.81
			1	211	0.81	1	0.9	
	Passive Aggressive	Set-1	0	53	0.87	0.87	0.87	0.88
			1	63	0.89	0.89	0.89	
		Set-2	0	63	0.93	0.86	0.89	0.9
			1	69	0.88	0.94	0.91	
		Set-3	0	61	0.9	0.77	0.83	0.88
			1	104	0.88	0.95	0.91	
		Set-4	0	53	0.92	0.64	0.76	0.92
			1	211	0.92	0.99	0.95	
	Support Vector Machine	Set-1	0	53	0.84	0.87	0.85	0.86
			1	63	0.89	0.86	0.87	
		Set-2	0	63	0.93	0.83	0.87	0.89
			1	69	0.86	0.94	0.9	
		Set-3	0	61	0.96	0.74	0.83	0.89
			1	104	0.86	0.98	0.92	
		Set-4	0	53	0.89	0.62	0.73	0.91
			1	211	0.91	0.98	0.95	
	Logistic Regression	Set-1	0	53	0.84	0.87	0.85	0.86
			1	63	0.89	0.86	0.87	
		Set-2	0	63	0.95	0.83	0.88	0.89
			1	69	0.86	0.96	0.90	
		Set-3	0	61	0.90	0.77	0.83	0.88
			1	104	0.88	0.95	0.91	
		Set-4	0	53	0.91	0.6	0.73	0.91
			1	211	0.91	0.99	0.95	

with logistic regression using count vectorizer also shown best performance on accuracy as we can see in Table 3 but the poor result on other three-parameter. The next highest accuracy of 89% is given by set 2 and set 3. Where set 2 given higher precision (90%), recall (86%), and f1-score (88%) at detecting fake news. In last we have our final result of support vector machine using TF-IDF given in Table 2. In here also set 4 shown the highest accuracy (91%). But poor performance on the other three-parameter. Set 2 and set 3 shown better performance as we can see in Table 2. Set 2 shown a precision of 93%, recall of 83%, f1-score of 87%, and accuracy of 89% which is combined best the other three sets. In Table 3, support vector machine using count vectorizer, as usual like other three algorithm support vector machine has shown highest accuracy of 90% in set 4 but the same poor result on other three-parameter. Set 3 shown the second highest accuracy of 88% and 84% on recall, precision, and f1-score. Set 2 gives an accuracy of 87%, the third-highest among four algorithms but having the higher value of precision (86%), recall (87%), and f1-score (87%) the other three algorithms.

Using set 4, almost all algorithms have given the highest accuracy except Multinomial Naïve bayes using TF-IDF. In all algorithms either using TF-IDF or count vectorizer method set 4 have given better accuracy but poor performance on other three-parameter. Set 1 and set 2 given decent and stable values. But set 2 was better in all algorithms either using TF-IDF or the count vectorizer method. So, from our observation, we can say that an excessive amount of one type of data can be misleading while detecting fake news. For our further analysis, we have taken set 2, as our final comparison shown in Fig. 10. Using the TF-IDF method Multinomial Naïve Bayes does not show a satisfying result. But other three algorithms showed better outcomes Fig. 10b. On the other hand, using the count vectorizer method all of the four algorithms shown better outcomes as we can see in Fig. 10a

VI. Conclusion

Fake news is one of the alarming issues of the digital era and fake news detection can halt this issue. Through our work, we want to contribute to solve this issue by

Table 3: Fake News Detection Result Using Count Vectorizer

Technique	Algorithms		T/F	Support	Precision	Recall	F1-Score	Accuracy
Count Vectorizer	Multinomial naïve bias	Set-1	0	53	0.89	0.77	0.83	0.85
			1	63	0.83	0.92	0.87	
		Set-2	0	63	0.89	0.86	0.87	0.88
			1	69	0.89	0.86	0.87	
		Set-3	0	61	0.88	0.80	0.84	0.88
			1	104	0.89	0.93	0.91	
		Set-4	0	53	0.79	0.77	0.78	0.91
			1	211	0.94	0.95	0.95	
	Passive Aggressive	Set-1	0	53	0.89	0.89	0.89	0.90
			1	63	0.90	0.90	0.90	
		Set-2	0	63	0.89	0.86	0.87	0.88
			1	69	0.87	0.90	0.89	
		Set-3	0	61	0.84	0.84	0.84	0.88
			1	104	0.90	0.90	0.90	
		Set-4	0	53	0.80	0.68	0.73	0.9
			1	211	0.92	0.96	0.94	
	Support Vector Machine	Set-1	0	53	0.82	0.85	0.83	0.84
			1	63	0.87	0.84	0.85	
		Set-2	0	63	0.86	0.87	0.87	0.87
			1	69	0.88	0.87	0.88	
		Set-3	0	61	0.84	0.84	0.84	0.88
			1	104	0.90	0.90	0.90	
		Set-4	0	53	0.80	0.68	0.73	0.90
			1	211	0.92	0.96	0.94	
	Logistic Regression	Set-1	0	53	0.89	0.79	0.84	0.86
			1	63	0.84	0.92	0.88	
		Set-2	0	63	0.90	0.86	0.88	0.89
			1	69	0.88	0.91	0.89	
		Set-3	0	61	0.88	0.82	0.85	0.89
			1	104	0.90	0.93	0.92	
		Set-4	0	53	0.92	0.62	0.74	0.91
			1	211	0.91	0.99	0.95	

presenting a data set containing more than 2900 fake and true news on Covid-19. In this digital era, people are mostly dependent on online news sources than printed news. As there are lots of news sources available online. So, there is a high risk of getting fake news on the online platform and fake news detection using the machine learning techniques can reduce the immense spread. To boost solving this issue, we have collected a new data set from traditional media based on Covid-19. We hope that This data set will help future researchers to contribute more to detecting fake news. We have also analyzed our data by four different existing supervised algorithms. We have calculated accuracy, recall, f1-score, and precision regarding count vectorizer and TF-IDF to detect fake news using algorithms and our data set on Covid-19. We have analyzed the algorithm with the four different amount of data. We have also analyzed how excessive amounts of one type of data can severely affect detection systems. As in this work, we only focused on the news related to Covid-19 and the news are from the traditional media so for this the number of data decreased. A largescale data set can be more effective in building a fake news detection system. As the fake news collection is challenging itself, especially from traditional media as there is a little amount of fake news in fact-checking sites. Also, we have faced similar kinds of fake news on different fact-checking sites. In the year 2020, many more events happened besides Covid-19. So, a positive path is to build a robust and large-scale data set of fake news from different media like social media, which researchers will use to facilitate additional study in this field. Using unsupervised or semi-supervised learning while detecting fake news can show better performance.

References Références Referencias

1. Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter, and Mohammad Rezwanul Huq. Detecting fake news using machine learning and deep learning algorithms. In 2019 7th International Conference on Smart Computing Communications (ICSCC), pages 1–5, 2019.
2. Vasu Agarwal, H. Parveen Sultana, Srijan Malhotra, and Amitrajit Sarkar. Analysis of classifiers for fake news detection. Procedia Computer Science, 165:377 – 383, 2019.
3. Sarah A. Alkhodair, Steven H.H. Ding, Benjamin C.M. Fung, and Junqiang Liu. Detecting breaking news rumors of emerging topics in social media.

Information Processing and Management, 57(2):102018, 2020.

4. Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2):211–236, May 2017.
5. Amjad Atallah. Iran: Over 700 dead after drinking alcohol to cure coronavirus, 2020. https://www.aljazeera.com/news/2020/4/27/iran-over-700-dead-after-drinking-alcohol-to-cure-coronavirus accessed August 26, 2020.
6. Pritika Bahad, Preeti Saxena, and Raj Kamal. Fake news detection using bi-directional lstm-recurrent neural network. Procedia Computer Science, 165:74–82, 2019. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC - DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.
7. Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation and cynicism. Communication Research, 41, 07 2012.
8. Paul Brewer, Dannagal Young, and Michelle Morreale. The impact of real news about "fake news": Intertextual processes and political satire. International Journal of Public Opinion Research, 25:323–343, 09 2013.
9. Jason Brownlee. What is a Confusion Matrix in Machine Learning, 2020. https://machinelearningmastery.com/confusion-matrix-machine-learning/ accessed August 22, 2020.
10. Alistair Coleman. 'Hundreds dead' because of Covid-19 misinformation, 2020. https://www.bbc.com/news/world-53755067 accessed August 22, 2020.
11. Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4, 2016.
12. Corinna Cortes and Vladimir VAPNIK. Support vector networks. Machine Learning, 20:273–297, 09 1995.
13. Shevon Desai. "Fake News," Lies and Propaganda: How to Sort Fact from Fiction, 2020. https://guides.lib.umich.edu/c.php?g=637508&p=4462356 accessed December 05, 2020.
14. Oberiri Destiny Apuke and Bahiyah Omar. Fake news and covid-19: modelling the predictors of fake news sharing among social media users. Telematics and Informatics, 2020.
15. Alan Duke. Lead Stories, 2020. https://leadstories.com/ accessed June 26, 2020.
16. Ahmet Dursun. Iran: Death toll from toxic alcohol rises to 180, 2020. https://www.aa.com.tr/en/health/iran-death-toll-from-toxic-alcohol-rises-to-180/1771659 accessed August 26, 2020.
17. Mykhailo Granik and Volodymyr Mesyura. Fake news detection using naive bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pages 900–903, 2017.
18. Mikell P. Groover. automation. Encyclopedia Britannica., 2020. https://www.brit annica.com/technology/autom ation accessed Dec ember 28, 2020.
19. Nolan Higdon. The anatomy of fake news : a critical news literacy education. University of California Press, 2020. Accessed November 28, 2020.
20. Angie Drobnic Holan. Politifact, 2020. https://www.politifact.com/ accessed June 30, 2020.
21. Junaed Younus Khan, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4:100032, 2021.
22. Lauren Lutzke, Caitlin Drummond, Paul Slovic, and Joseph Arvai. Priming critical thinking: Simple interventions limit the influence of fake news about climate change on facebook. Global Environmental Change, 58:101964, 09 2019.
23. Andrew Mccallum and Kamal Nigam. A comparison of event models for naive bayes text classification. Work Learn Text Categ, 752, 05 2001.
24. Claire Milne. Full Fact, 2020. https://fullfact.org/ accessed August 10, 2020.
25. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
26. Martin Porter. An algoritm for suffix stripping. Program: electronic library and information systems, 40:211–218, 07 2006.
27. Anand Rajaraman and Jeffrey David Ullman. Data Mining. Cambridge University Press, 2011.
28. Lori Robertson. Fact Check, 2020. https://www.factcheck. org/(accessed June 30, 2020).
29. Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A Hybrid Deep Model for Fake News Detection, page 797–806. Association for Computing Machinery, New York, NY, USA, 2017.
30. Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern Approach. Prentice Hall Press, USA, 3rd edition, 2009.
31. Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. A comparative analysis of logistic regression, random forest and knn models for the text classification. Augmented Human Research, 5, 12 2020.
32. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social

media: A data mining perspective. SIGKDD Explor. Newsl., 19(1):22–36, September 2017.

33. Maciej Szpakowski. FakeNewsCorpus, 2020. https://github.com/several27/FakeNewsCorpus accessed November 23, 2020.
34. Mark Trudeau. COVID-19 IMPACT: Broadband usage surges 47% IN Q1, nearing YE2020 expectations, with exponential rise in >1TB/2TB power users, 2020. https://openva ult.com/com plimentary-report-Q12/accessed Aug ust 22, 2020.
35. Emmanuel Vincent. Health Feedback, 2020. https:// hea lthfeed back.org/accessed July 04, 2020.
36. William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection". Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, page 422–426, 2017.
37. Fan Yang, Arjun Mukherjee, and Eduard Dragut. Satirical news detection and analysis using attention mechanism and linguistic features. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1979–1989. Association for Computational Linguistics, sep 2017.
38. Chaowei Zhang, Ashish Gupta, Christian Kauten, Amit V. Deokar, and Xiao Qin. Detecting fake news for reducing misinformation risks using analytics approaches. European Journal of Operational Research, 279(3):1036–1052, 2019.
39. Xichen Zhang and Ali A. Ghorbani. An overview of online fake news: Characterization, detection, and discussion. Information Processing and Management, 57(2):102025, 2020.