PURPOSE-LED
PUBLISHING™

PAPER • OPEN ACCESS

Fake News Detection Using Machine Learning Approaches

To cite this article: Z Khanam *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1099** 012040

View the article online for updates and enhancements.

Fake News Detection Using Machine Learning Approaches

Z Khanam[1], B N Alwasel[1], H Sirafi[1] and M Rashid[2]

[1]College of Computing and Informatics, Saudi Electronic University, Dammam, KSA
[2]School of Computer Science and Engineering, Lovely Professional University, Jalandhar, India

 E-mail: mamoon873@gmail.com

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

1. Introduction

Fake News contains misleading information that could be checked. This maintains lie about a certain statistic in a country or exaggerated cost of certain services for a country, which may arise unrest for some countries like in Arabic spring. There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues as confirming authors are accountable. However, their scope is so limited because they depend on human manual detection, in a globe with millions of articles either removed or being published every minute, this cannot be accountable or feasible manually. A solution could be, by the development of a system to provide a credible automated index scoring, or rating for credibility of different publishers, and news context.

This paper proposes a methodology to create a model that will detect if an article is authentic or fake based on its words, phrases, sources and titles, by applying supervised machine learning algorithms on an annotated (labeled) dataset, that are manually classified and guaranteed. Then, feature selection methods are applied to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results. We propose to create the model using different classification algorithms. The product model will test the unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and can be used and integrated with any system for future use.

2. Related Work

2.1 Social Media and Fake News

Social media includes websites and programs that are devoted to forums, social websites, microblogging, social bookmarking and wikis [1][2]. On the other side, some researchers consider the fake news as a result of accidental issues such as educational shock or unwitting actions like what happened in Nepal Earthquake case [3][4]. In 2020, there was widespread fake news concerning health that had exposed global health at risk. The WHO released a warning during early February 2020 that the COVID-19 outbreak has caused massive 'infodemic', or a spurt of real and fake news—which included lots of misinformation.

2.2 Natural Language Processing

The main reason for utilizing Natural Language Processing is to consider one or more specializations of system or an algorithm. The Natural Language Processing (NLP) rating of an algorithmic system enables the combination of speech understanding and speech generation. In addition, it could be utilized to detect actions with various languages.[6] suggested a new ideal system for extraction actions from languages of English, Italian and Dutch speeches through utilizing various pipelines of various languages such as Emotion Analyzer and Detection, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Chunking, and Semantic Role Labeling made NLP good Subject of the search [5][6].

The Sentiment analysis [7] extracts emotions on a particular subject. Sentiment analysis is composed of extracting a specific term for a subject, extracting the sentiment, and pairing with connection analysis. The Sentiment analysis uses dual languages Resources for analysis: Glossary of meaning and Sentiment models database. for constructive and Destructive words and attempts to give classifications on a level of -5 to 5. Parts of speech taggers tools for languages such as European languages are being explored to produce parts of language taggers tools in different languages such as Sanskrit [8], Hindi [9] and Arabic. Can be efficient Mark and categorize words as names, adjectives, verbs, and so on. Most part of speech techniques can be performed effectively in European languages, but not in Asian or Arabic languages. Part of the Sanskrit word "speak" specifically uses the tree-bank method. The Arabic utilizes Vector Machine (SVM) [10] uses a method to automatically identify symbols and parts of speech and automatically expose basic sentences in Arabic text [11].

2.3 Data Mining

Data mining techniques are categorized into two main methods, which is; supervised and unsupervised. The supervised method utilizes the training information in order to foresee the hidden activities. Unsupervised Data Mining is a try to recognize hidden data models provided without providing training data for example, pairs of input labels and categories. A model example for unsupervised data mining is aggregate mines and a syndicate base [12].

2.4 Machine Learning (ML) Classification

Machine Learning (ML) is a class of algorithms that help software systems achieve more accurate results without having to reprogram them directly. Data scientists characterize changes or characteristics that the model needs to analyze and utilize to develop predictions. When the training is completed, the algorithm splits the learned levels into new data [11]. There are six algorithms that are adopted in this paper for classifying the fake news.

2.5 Decision Tree

The decision tree is an important tool that works based on flow chart like structure that is mainly used for classification problems. Each internal node of the decision tree specifies a condition or a "test" on an attribute and the branching is done on the basis of the test conditions and result. Finally the leaf node bears a class label that is obtained after computing all attributes. The distance from the root to leaf represents the classification rule. The amazing thing is that it can work with category and dependent variable. They are good in identifying the most important variables and they also depict the relation

between the variables quite aptly. They are significant in creating new variables and features which is useful for data exploration and predicts the target variable quite efficiently.

Decision Tree Pseudo-code
GenerateDecisionTree(Sample s, features F)
1. If stop _conditions(S,F) = true then
    a.   leaf = create_Node()
    b.   Leaf.lable= classify(s)
    c.   Return leaf
2. root = create_Node()
3. root.testcondition = find_bestSplit(s,f)
4. v = { v l v a possible outcome of root.testconditions)
5. for each value v∈ V:
6. sv: = {s │ root.testcondition(s) = v and s∈ S};
7. child = Tree_Growth(Sv ,F) ;
8. Grow child as a descent of roof and label the edge (root→child) as v
        Return root

Tree based learning algorithms are widely with predictive models using supervised learning methods to establish high accuracy. They are good in mapping non-linear relationships. They solve the classification or regression problems quite well and are also referred to as CART [13][14][15].

2.6 Random Forest
 Random Forest are built on the concept of building many decision tree algorithms, after which the decision trees get a separate result. The results, which are predicted by large number of decision tree, are taken up by the random forest. To ensure a variation of the decision trees, the random forest randomly selects a subcategory of properties from each group [16][17]
The applicability of Random forest is best when used on uncorrelated decision trees. If applied on similar trees, the overall result will be more or less similar to a single decision tree. Uncorrelated decision trees can be obtained by bootstrapping and feature randomness.

Random Forest Pseudo-code
To make *n* classifiers:
For *i* = 1 to *n* **do**
Sample the training data T randomly with replacement for Ti output
Build a Ti-containing root node, Ni
Call BuildTree (*Ni*)
end For
BuildTree (N):
If N includes instances of only one class, then returns
else
    Select z% of the possible splitting characteristics at random in N
    Select the feature *F* with the highest information gain to split on
    Create f child nodes of *N, Ni ,..., Nf* , where *F* has *f* possible values (*F*1, ... , *Ff*)
For *i* = 1 to *f* **do**
        Set the contents of *Ni* to *Ti*, where *Ti* is all instances in *N* that match *Fi*
        Call Buildtree (*Ni*)
end for
end if [18]

2.7   Support Vector Machine (SVM)

The SVM algorithm is based on the layout of each data item in the form of a point in a range of dimensions n (the number of available properties), and the value of a given property is the number of specified coordinates [13]. Given a set of n features, SVM algorithm uses n dimensional space to plot the data item with the coordinates representing the value of each feature. The hyper-plane obtained to separate the two classes is used for classifying the data.

SVM Pseudo-Code

F[0..N-1]: a feature set with N features that is sorted by information gain in decreasing order accuracy(i): accuracy of a prediction model based on SVM with F[0...i] gone set

```
  low = 0
  high = N-1
  value = accuracy(N-1)
  IG_RFE_SVM(F[0...N-1], value, low, high) {
      If (high }≤ low)
        Return F[0...N-1] and value
    mid = (low + high ) / 2
    value_2 = accuracy(mid)
    if (value_2 ≥  value)
        return IG_RFE_SVM(F[0...mid], value_2, low, mid)
    else (value_2 < value)
         return IG_REF_SVM(F[0...high], value, mid, high) [13]
```

2.8. Naive Bayes

This algorithm works on Bayes theory under the assuming that its free from predictors and is used in multiple machine learning problems [18]. Simply put, Naive Bayes assumes that one function in the category has nothing to do with another. For example, the fruit will be classified as an apple when its of red color, swirls, and the diameter is close to 3 inches. Regardless of whether these functions depend on each other or on different functions, and even if these functions depend on each other or on other functions, Naive Bayes assumes that all these functions share a separate proof of the apples [14]

Naive Bayes Equation

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x1 \mid c) \times P(x2 \mid c) \times \ldots \times P(x2 \mid c) \times P(c)$$

Where:

$P(c \mid X)$ is the posterior Probability.

$P(x \mid c)$ is the Likelihood.

$P(c)$ is the Class Prior Probability.

$P(x)$ is the Predictor Prior Probability.

Naive Bayes Pseudo-code

Training dataset T,

F= (f1, f2, f3,..., fn)   // value of the predictor variable in testing dataset.

Output:

A class of testing dataset.

Step:

1.   Read Training Dataset T;
2.   Calculate the mean and norm of each class's predictor variables;
3.   Repeat
4.   Calculating the likelihood of using the equation of gauss density in each class;
5.   Until Pending the estimation of the likelihood of all predictor variables (f1, f2, f3,..., fn),.
6.   Calculated the likelihood for respective class;

7.   Get the highest likelihood;
8.   (Researchgate.net, 2018)

Random Forest (RF) and Naïve Bayes have many differences, the main is their model size. The NB models are not good at representing complex behavior, resulting in low model size and good for a constant type of data. In contrast, the model sixe for Random Forest model is very large and it might results in over fitting. NB is good for dynamic data and can be reshaped easily when new data is inserted while using a RF may require a rebuild of the forest every time a change is introduced.

2.9. KNN (k- Nearest Neighbors)
KNN classifies new positions based on most of the sounds from the neighboring k with respect to them. The position assigned in the class is highly mutually exclusive between the nearest neighbors K, as measured by the role of the distance [15].

> ### KNN Pseudo-code
> Classify ($\mathbf{X}$, $\mathbf{Y}$, x) // $\mathbf{X}$: training data, $\mathbf{Y}$: class labels of $\mathbf{X}$, x: unidentified sample
> **For** $i = 1$ to m **do**
>   Calculate distance d ($\mathbf{X}i$, x)
> **end for**
> Calculate set I containing indices for k smallest distances d ($\mathbf{X}i$, x).
> **return** majority label for {$\mathbf{Y}i$ where $i \in I$}[16]

 KNN falls in the category of supervised learning and its main applications are intrusion detection, pattern recognition. It is nonparametric, so no specific distribution is assigned to the data or any assumption is made about them. For example GMM, assumes a Gaussian distribution of the given data.

2.10. Combining Classifiers
Achieving the best possible taxonomic performance is the primary goal when planning paradigm-detecting systems. For that reason, different classification planners for the models of detecting actions are able to be progressed. Although if one model may perform the highest execution, the style sets correctly categorized by variant classifiers is not important to be overlap. Variant categorization planners can give additional information for the models. With this additional information, the execution of individual models can be improved [19].

2.11 Related Work on Fake News Detection
[20] pointed out various sources of media and made the suitable studies whether the submitted article is reliable or fake. The paper utilizes models based on speech characteristics and predictive models that do not fit with the other current models.

[21] used naïve Bayes classifier to detect fake news by Naive Bayes. This method was performed as a software framework and experimented it with various records from the Facebook, etc., resulting in an accuracy of 74%. The paper neglected the punctuation errors, resulting in poor accuracy.
[22] estimated various ML algorithms and made the researches on the percentage of the prediction. The accuracy of various predictive patterns included bounded decision trees, gradient enhancement, and support vector machine were assorted. The patterns are estimated based on an unreliable probability threshold with 85-91% accuracy.
[23] utilized the Naive Bayes classifier, discuss how to implement fake news discovery to different social media sites. They used Facebook, Twitter and other social media applications as a data sources for news. Accuracy is very low because the information on this site is not 100% credible.
[24][25][26] discuss misleading and discovering rumors in real time. It utilizes a novelty-based characteristic and derives its data source from Kaggle. The accuracy average of this pattern is 74.5%. Clickbait and sources do not consider unreliable, resulting in a lower resolution.

[27] Used to distinguish Twitter spam senders. Among the various models used are the naive Bayes algorithms, the clustering, and the decision tree. The accuracy average of detecting spammers is 70% and fraudsters 71.2%. The models used have achieved a low level of intermediate precision to separate spammers from non-spam.[28] identified fake news in different ways. The accuracy is limited to 76% as a language model. Greater accuracy can be achieved if a predictive model is used.

[29] aimed to utilize machine learning methods to detect fake news. Three common methods are utilized through their researches: Naïve Bayes, Neural Network and Support Vector Machine (SVM). Normalization technique is an essential stage in data cleansing prior machine learning is used to categorizing the data. The output proved that that Naïve Bayes has an accuracy of 96.08% for detecting fake messages. Two more advanced methods, the neural network and the machine vector (SVM) reached an accuracy of 99.90%.

In [30] it has been discovered that fake news detection is a predictive analysis application. Detecting counterfeit messages involves the three stages of processing, feature extraction and classification. The hybrid classification model in this research is designed for Show fake news. The combination of classification is a combination of KNN and random forests. The execution of the suggested model is analyzed for accuracy and recall. the final results improved by up to 8% using a mixed false message detection model.

[31][32] examined how fake news was used in the 2012 Dutch elections on Twitter. She examines the execution of 8 supervised machine learning classifiers in the Twitter dataset [33]. We assume that the decision tree algorithm works best for the data set used with a F score of 88%. 613,033 tweets were rated, of which 328,897 were considered genuine and 284,136 were false. By analyzing the qualitative content of false tweets sent during the election, features and properties of the wrong content were found and divided into six different categories [34].

[30] presented a counterfeit detection model using N-gram analysis by the lenses of various characteristic extraction techniques. In addition, we examined the extraction techniques of various features and six different methods of machine learning. The proposed model achieves the highest accuracy in use Contains a unigram and a linear SVM workbook. The highest accuracy is 92%.

3.   Methodology

This section presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection phase, followed by preprocessing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers [35][36][37][38][39]. Figure [1] describes the proposed system methodology. The methodology is based on conducting various experiments on dataset using the algorithms described in the previous section named Random forest, SVM and Naïve Bayes, majority voting and other classifiers. The experiments are conducted individually on each algorithm, and on combination among them for the purpose of best accuracy and precision [40][41][42].
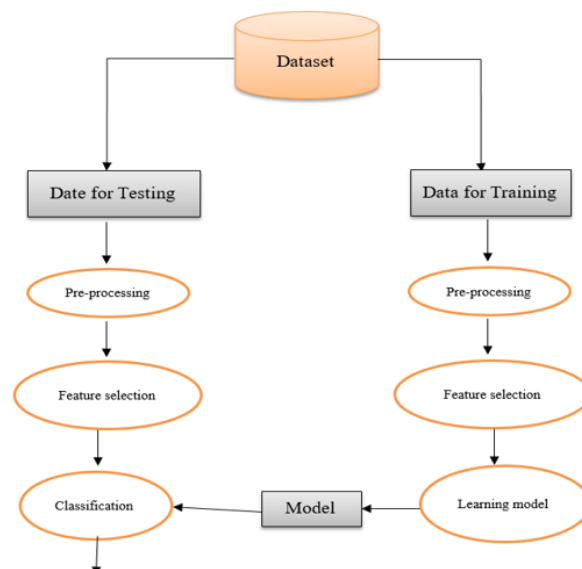
Figure 1. Describes the Proposed System Methodology

The main goal is to apply a set of classification algorithms to obtain a classification model in order to be used as a scanner for a fake news by details of news detection and embed the model in python application to be used as a discovery for the fake news data [43][44]. Also, appropriate refactorings have been performed on the Python code to produce an optimized code [25][26].

The classification algorithms applied in this model are k-Nearest Neighbors (k-NN), Linear Regression, XGBoost, Naive Bayes, Decision Tree, Random Forests and Support Vector Machine (SVM). All these algorithms get as accurate as possible. Where reliable from the combination of the average of them and compare them.

As shown in the figure [2], the dataset is applied to different algorithms in order to detect a fake news. The accuracy of the results obtained are analyzed to conclude the final result.
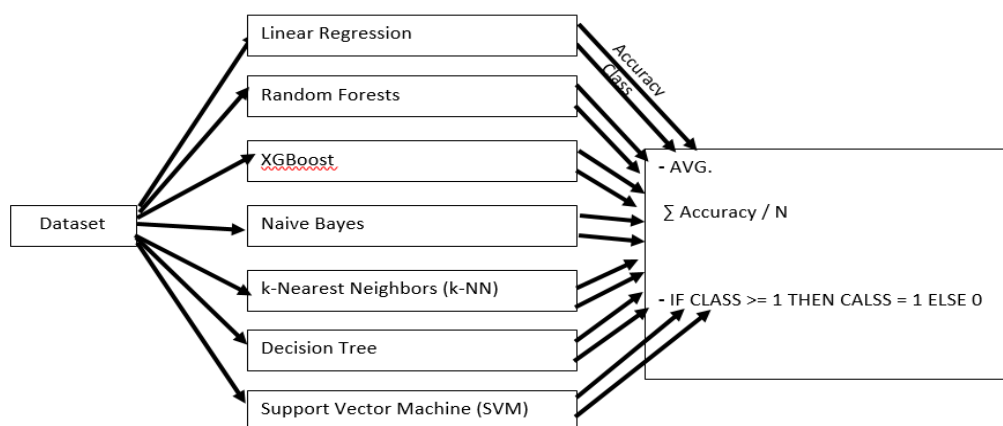

Figure 2. The Classification Algorithms

In the process of model creation, the approach to detecting political fake news is as follows: First step is collection political news dataset, (the Liar dataset is adopted for the model), perform preprocessing through rough noise removal, the next step is to apply the NLTK (Natural Language Toolkit) to perform POS and features are selected. Next perform the dataset splitting apply ML algorithms (Naïve bays and Random forest) then create the proposed classifier model. The Fig 2 shows that after the NLTK is applied, the Dataset gets successfully preprocessed in the system, then a message is generated for

applying algorithms on trained portion. The system response with N.B and Random forest are applied, then the model is created with response message. Testing is performed on test dataset, and the results are verified, the next step is to monitor the precision for acceptance. The model is then applied on unseen data selected by user. Full dataset is created with half of the data being fake and half with real articles, thus making the model's reset accuracy 50%. Random selection of 80% data is done from the fake and real dataset to be used in our complete dataset and leave the remaining 20% to be used as a testing set when our model is complete. Text data requires preprocessing before applying classifier on it, so we will clean noise, using Stanford NLP (Natural language processing) for POS (Part of Speech) processing and tokenization of words, then we must encode the resulted data as integers and floating point values to be accepted as an input to ML algorithms. This process will result in feature extraction and vectorization; the research using python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer and Tiff Vectorizer. Data is viewed in graphical presentation with confusion matrix. Refer figure 3.
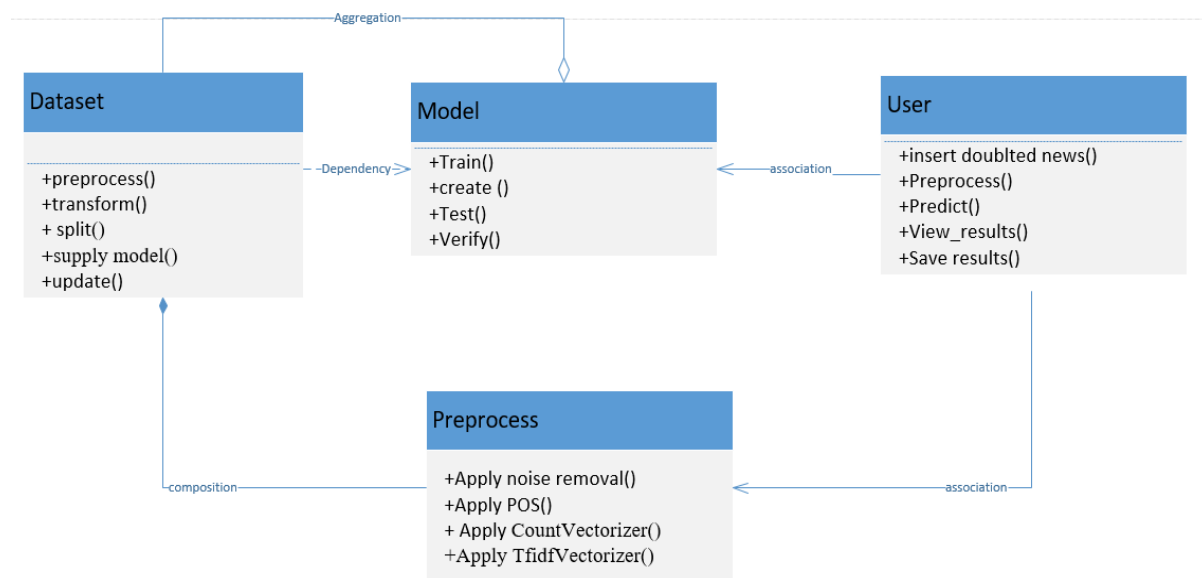

Figure 3. Fake Detector Model

This section discusses the chosen dataset, The LIAR-PLUS Master that has been used for cleaning and extracting the data set, and the algorithms are applied. This dataset has automatically extracted proof sentences from the full-text verdict article published in Politifact by journalists.

As shown in following figure 4 we used the features of Truth values, in addition we applied part of speech on the statement to get another 4 features (nouns, verbs, preposition and sentences) and each record is labeled by class label as (0, 1, 2, 3) to be used in training the model. The following steps have been used to evaluate the precision of the news.

1.   Liar-dataset is preprocessed 12.8K
2.   The texts in multiple contexts is taken from POLITIFACT.COM and is labelled manually. Then it is transformed from TSV format into CSV format using Python.
3.   The next step is to clean the noise using NLP NLTK libraries and SAFAR v2 library. The noise involves ids, dots, commas, quotations, and by stemming terms, delete the suffix. The next step is to use POS (Part of speech) which will turn the dataset into tokens and statistical values.
4.   Perform feature extraction by choosing lexical features, Such as word count, average word length, length of article, number count, number of sections of speech (adjective).

5. Extract unigram and bigram features by using Tfidf Vectorizer function of python sklearn. Feature extraction library to generate TF-IDF n-gram features.
6. Divide the dataset into 70% for train and 30% for test using python sklearn.
7. Produce classification model ipynb file after applying all the algorithms.
8. Test model precision on the test portion of dataset and produce confusion matrix.
9. Evaluate accuracy, precision, recall, and f1-score for fake and real news.
10. Design the interface to be used for testing unseen news by user.

	ID	label	statement	barely true counts	false counts	half true counts	mostly true counts	pants on fire counts
0	2635	0	Says the Annies List political group supports ...	0	1	0	0	0
1	10540	3	When did the decline of coal start? It started...	0	0	1	1	0
2	324	3	Hillary Clinton agrees with John McCain "by vo...	70	71	160	163	9
3	1123	0	Health care reform legislation is likely to ma...	7	19	3	5	44
4	9028	3	The economic turnaround started at the end of ...	15	9	20	19	2
5	12465	1	The Chicago Bears have had more starting quart...	0	3	2	5	1
6	2342	3	Jim Dunnam has not lived in the district he re...	3	1	1	3	1
7	153	3	I'm the only person on this stage who has work...	70	71	160	163	9
8	5602	3	However, it took $19.5 million in Oregon Lotte...	0	0	1	0	1
9	9741	3	Says GOP primary opponents Glenn Grothman and	0	0	0	1	0

	ID	label	statement	barely true counts	false counts	half true counts	mostly true counts	pants on fire counts	Nouns	Verbs	Preposition	Sentences
0	12465	1	The Chicago Bears have had more starting quart...	0	3	2	5	1	8	3	4	1
1	620	1	McCain opposed a requirement that the governme...	70	71	160	163	9	6	4	1	2
2	13237	1	Says Paul Ryan is still endorsing Trump.	40	29	69	78	7	3	3	0	1
3	7550	1	We have a federal government that thinks they ...	36	33	15	19	8	4	4	1	1
4	9018	1	Austin is a city that has basically doubled in...	1	0	4	1	0	4	5	2	1
5	1038	1	The nuclear test conducted in our nation this ...	0	0	0	0	0	11	4	3	2
6	8492	1	Under President George W. Bush, ...	2	1	1	4	0	9	2	2	2

Figure 4. Dataset LIAR after Preprocessing

The data is divided it into two parts: The first section, which consists of **75%** of the data, is a trained data, where the algorithm detects the real news and false news ,then the data is labeled in the form of 0 and 1 where 0 is for false news and 1 for true news. After that, the rest of the data, which is **25%** of it, will do a test on it, so that it is sure whether the news is nature or forged, and then return it in case it was right or wrong, and according to the percentage of right and wrong, the algorithm percentage will be formed. Refer figure 5 and 6.

```
In [10]: df = pd.read_csv('output01.csv', encoding="utf_8")
         df
```

			$800,000 from t...									
26	3054	1	Whether you like to admit it or not, half our ...	0	3	1	1	0	2	3	1	1
27	9554	1	There are already more American jobs in the so...	2	3	2	7	0	5	1	3	1
28	2537	1	Webster sponsored a bill to create a form of m...	6	5	4	11	1	6	3	2	1
29	4681	1	Says judges get better benefits at a lower cos...	10	17	27	19	8	5	2	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...
3048	5647	0	Says new estimates from the Congressional Budg...	0	1	0	0	0	8	5	3	1
3049	5123	0	Mitt Romney has 15 homes.	2	2	3	1	0	3	1	0	1
3050	2960	0	[State Rep. Doug McKillips] donors deserve a r...	0	1	0	1	0	7	1	0	1
3051	8591	0	On changing the rules for filibusters on presi...	7	6	5	7	0	3	1	3	1

Figure 5. Dataset LIAR after Preprocessing

```
In [13]:  from sklearn.model_selection import train_test_split
          X = df[['barely true counts','false counts', 'half true counts','mostly true counts','pants on fire counts','Nouns','Ve
          y = df['label']
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=1)

In [75]:  # X_train

In [76]:  # X_test

In [13]:  # y_train

In [14]:  # y_test
```

Figure 6. Dataset LIAR after Preprocessing

4. Results

The scope of this project is to cover the political news data, of a dataset known as Liar-dataset, it is a New Benchmark Dataset for Fake News Detection and labeled by fake or trust news. We have performed analysis on "Liar" dataset . The results of the analysis of the datasets using the six algorithms have been depicted using the confusion matrix. The six algorithms used for the detection are as:

- XGboost.
- Random Forests.
- Naive Bayes.
- K-Nearest Neighbors (KNN).
- Decision Tree.
- SVM

The confusion matrix is automatically obtained by Python code using the cognitive learning library when running the algorithm code in Anaconda platform.

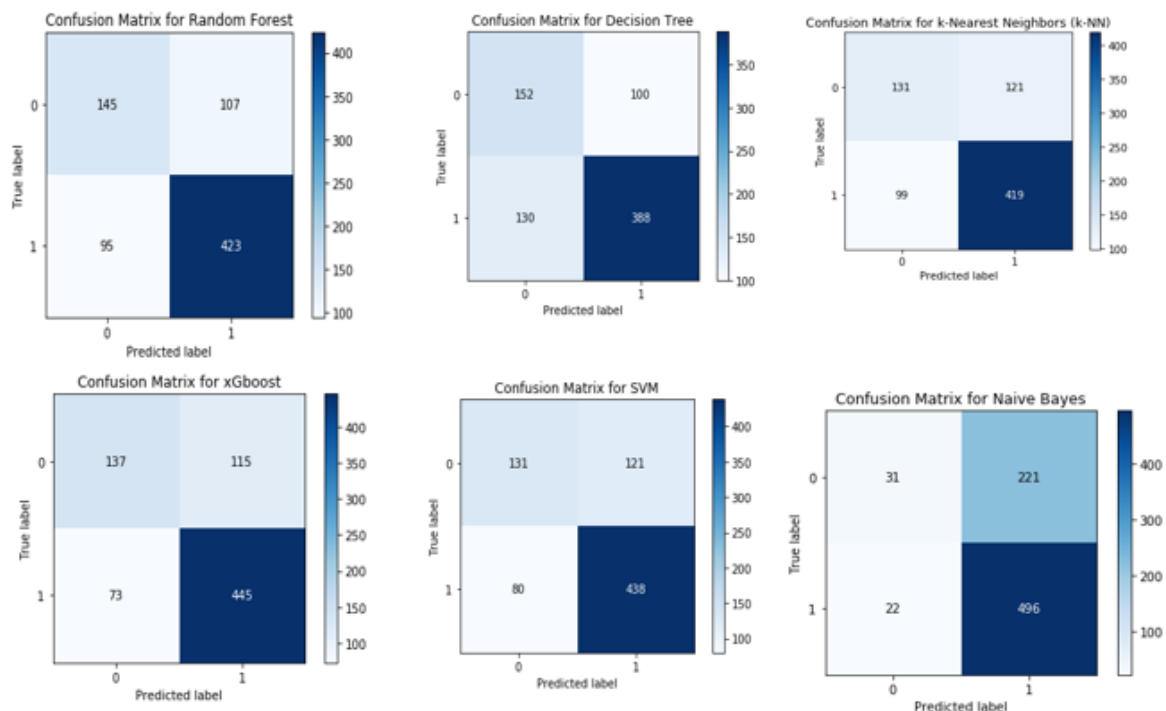The Confusion Matrix for all the algorithms are depicted below in figure 7:

Figure 7. Confusion Matrix Results for the Algorithms

Figure 8 expresses the accuracies of these algorithms. As shown the XGBOOST is depicting the highest accuracy with more than 75%, next is SVM and Random forest with approximately 73% accuracy.
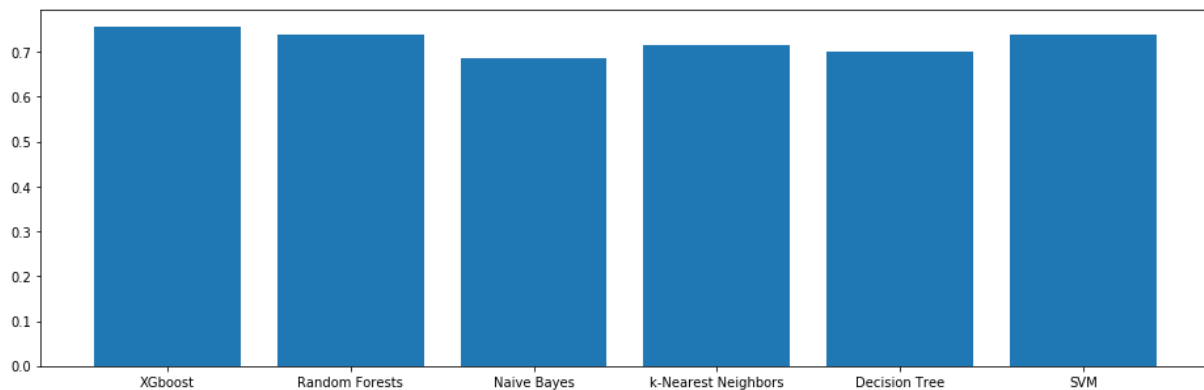
Figure 8. Accuracy Results of all the Algorithms

5. Conclusion

The research in this paper focuses on detecting the fake news by reviewing it in two stages: characterization and disclosure. In the first stage, the basic concepts and principles of fake news are highlighted in social media. During the discovery stage, the current methods are reviewed for detection of fake news using different supervised learning algorithms.

As for [20] the displayed fake news detection approaches that is based on text analysis in the paper utilizes models based on speech characteristics and predictive models that do not fit with the other current models. From [21] they utilized Naive Bayes classifier to detect fake news from different sources, with results of accuracy of 74%. [22] Used combined ML algorithms, but they depend on unreliable probability threshold with 85-91% accuracy. [23] uses the Naive Bayes to detect fake news from different social media websites, but the results were not accurate for the untruthful sources. [24] They got their data from Kaggle with average accuracy of 74.5%. [27] Used Naive Bayes algorithms to detect Twitter spam senders, with accuracy rated from 70% to 71.2%. [28] They tried different approaches with accuracy of 76%. [29] Three common methods are utilized through their researches: Naïve Bayes, Neural Network and Support Vector Machine (SVM). Naïve Bayes has an accuracy of 96.08% for detecting fake messages. The neural network and the machine vector (SVM) reached an accuracy of 99.9 0%. [30] They used the combination of KNN and random forests that gave the final results improved by up to 8% using a mixed false message detection model. [31] They worked on 2012 Dutch elections fake news on Twitter, they examine the execution of 8 supervised machine learning classifiers in the Twitter dataset. And they assume that the decision tree algorithm works best for the data set used with a F score of 88%. [32] Presented a counterfeit detection model using N-gram analysis achieved the highest accuracy in use contains a unigram and a linear SVM workbook. The highest accuracy is 92%.

In the aforementioned research summary and system analysis, we concluded that most of the research papers used naïve bays algorithm, and the prediction precision was between 70-76%, they mostly use qualitative analysis depending on sentiment analysis, titles, word frequency repetition [40][41][42]. In our approach we propose to add to these methodologies, another aspect, which is POS textual analysis, it is a quantitative approach, it depends on adding numeric statistical values as features, we thought that increasing these features and using random forest will give further improvements to precession results. The features we propose to add in our dataset are total words (tokens), Total unique words (types), Type/Token Ratio (TTR), Number of sentences, Average sentence length (ASL), Number of characters, Average word length (AWL), nouns, prepositions, adjectives etc.

6. References

[1].   Economic and Social Research Council. Using Social Mmedia. Available at: https://esrc.ukri.org/research/impact-toolkit/social-media/using-social-media
[2].   Gil, P. Available at: https://www.lifewire.com/what-exactly-is-twitter-2483331. 2019, April 22.

[3]. E. C. Tandoc Jr et al. "Defining fake news a typology of scholarly definitions". *Digital Journalism* , 1–17. 2017.

[4]. J. Radianti et al. "An Overview of Public Concerns During the Recovery Period after a Major Earthquake: Nepal Twitter Analysis." *HICSS '16 Proceedings of the 2016 49th Hawaii International Conference on System Sciences* (HICSS) (pp. 136-145). Washington, DC, USA : IEEE. 2016.

[5]. Alkhodair S A, Ding S H.H, Fung B C M, Liu J 2020 "Detecting breaking news rumors of emerging topics in social media" *Inf. Process. Manag.* 2020, 57, 102018.

[6]. Jeonghee Yi et al. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. "In Data Mining, 2003. ICDM 2003. *Third IEEE International Conference* (pp. 427-434). http://citeseerx.ist.psu.edu. 200).2003

[7]. Tapaswi et al. "Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit   m   sentences." *Software Engineering (CONSEG), on Software Engineering* (CONSEG), (pp. 1-4). IEEE. 2012

[8]. Ranjan et al. "Part of speech tagging and local word grouping techniques for natural language parsing in Hindi". *In Proceedings of the 1st International Conference on Natural Language Processing* (ICON 2003).   Semanticscholar. 2003

[9]. MonaDiab et al. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. *Proceedings of HLT-NAACL 2004: Short Papers* (pp. 149–152). Boston, Massachusetts, USA: Association for Computational Linguistics. 2004

[10]. Rouse, M. https://searchenterpriseai.techtarget.com/definition/machine-learning-ML May 2018

[11]. Sumeet Dua, Xian Du. "Data Mining and Machine Learning in Cybersecurity". *New York: Auerbach   Publications.*19 April 2016.

[12]. RAY, S. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/ 2017, September

[13]. Huang, T.-Q. (n.d.) https://www.researchgate.net/figure/Pseudo-code-of-information-gain-based-recursive-feature-elimination-procedure-with-SVM_fig2_228366941 2018

[14]. Researchgate.net. Available at: Available at: https://www.researchgate.net/figure/Pseudocode-of-naive-bayes-algorithm_fig2_325937073. 2018.

[15]. Researchgate.net.   Available   at: https://www.researchgate.net/figure/Pseudocode-for-KNN-classification_fig7_260397165, 2014.

[16]. Rampersad G, Althiyabi T 2020 "Fake news: Acceptance by demographics and culture on social media" *J. Inf. Technol. Politics* 2020, 17, 1–11.

[17]. NaphapornSirikulviriya; SukreeSinthupinyo. "Integration of Rules from a Random Forest." *International Conference on Information and Electronics Engineering* (p. 194 : 198). Singapore: semanticscholar.org. 2011.

[18]. Jasmin Kevric et el. "An effective combining classifier approach using tree algorithms for network intrusion detection." *Neural Computing and Applications* , 1051–1058. 2017.

[19]. ShivamB.Parikh and PradeepK.Atrey. "Media-RichFake News Detection: A Survey." *IEEE Conference on Multimedia Information*. Miami, FL: IEEE. 2018.

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

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

[22]. Akshay Jain and AmeyKasbe. "Fake News Detection." *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science* (SCEECS). Bhopal, India: IEEE. 2018.

[23]. Yumeng Qin et al. "Predicting Future Rumours." *Chinese Journal of Electronics* (Volume: 27 , Issue: 3 , 5 2018, 514 - 520.

[24]. ArushiGupta and RishabhKaushal. "Improving spam detection in Online Social Networks." *International Conference on Cognitive Computing and Information Processing* (CCIP). semanticscholar.org.2015.

[25]. Khanam, Z., Ahsan, M.N."Evaluating the effectiveness of test driven development: advantages and   pitfalls."*International. J. Appl. Eng. Res*. 12, 7705–7716, 2017

[26]. Khanam, Z. "Analyzing refactoring trends and practices in the software industry." *Int. J. Adv. Res. Comput. Sci*. 10, 0976–5697, 2018.

[27]. Veronica   Perez-Rosas   et   al.   Available   at: https://www.researchgate.net/publication/319255985_Automatic_Detection_of_Fake_News August, 2017.

[28]. Supanya Aphiwongsophon et al. " Detecting Fake News with Machine Learning Method." 2018 *15th   International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (ECTI-CON). Chiang Rai, Thailand, Thailand: IEEE . 2018.

[29]. Prabhjot Kaur et al. " Hybrid Text Classification Method for Fake News Detection." *International Journal of Engineering and Advanced Technology* (IJEAT) , 2388-2392. 2019.

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

[31]. I. Traore et al. "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques." *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments* (pp. 127–138). Springer International Publishing . 2017.

[32]. Khanam Z., Alkhaldi S. "An Intelligent Recommendation Engine for Selecting the University for Graduate Courses in KSA: SARS Student Admission Recommender System." *In: Smys S., Bestak R., Rocha Á. (eds) Inventive Computation Technologies*. ICICIT 2019. Lecture Notes in Networks and Systems, vol 98. Springer, Cham. 2019.

[33]. Khanam Z. and Ahsan M.N. " Implementation of the pHash algorithm for face recognition in secured remote online examination system." *International Journal of Advances in Scientific Research and Engineering* (ijasre) Volume 4, Issue 11 November. 2018.

[34]. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology* (TIST), 10(3), 1-42.

[35]. Sharma, Karishma, et al. "Combating fake news: A survey on identification and mitigation techniques." *ACM Transactions on Intelligent Systems and Technology* (TIST) 10.3 (2019): 1-42.

[36]. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1), 22-36.

[37]. Shu, Kai, et al. "Fake news detection on social media: A data mining perspective." *ACM SIGKDD explorations newsletter* 19.1 (2017): 22-36.

[38]. Khanam Z. and Agarwal S. Map-reduce implementations: Survey and Performance comparison, *International Journal of Computer Science & Information Technology* (IJCSIT) Vol 7, No 4, August 2015.

[39]. Zhang, Jiawei, Bowen Dong, and S. Yu Philip. "Fakedetector: Effective fake news detection with deep diffusive neural network." *2020 IEEE 36th International Conference on Data Engineering* (ICDE). IEEE, 2020.

[40]. K Ludwig, M Creation 2020 "Dissemination and uptake of fake-quotes in lay political discourse on Facebook and Twitter" *J. Pragmat*, 157, 101–118.

[41]. Can Machines Learn to Detect Fake News? A Survey Focused on Social Media.   Available at: https://scholarspace.manoa.hawaii.edu/handle/10125/59713

[42]. Cardoso Durier da Silva, F., Vieira, R., & Garcia, A. C. (2019, January). Can machines learn to detect fake   news? a survey focused on social media. *In Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[43]. Bovet, Alexandre, and Hernán A. Makse. "Influence of fake news in Twitter during the 2016 US presidential election." *Nature communications* 10.1 (2019): 1-14. The science of fake news.

[44]. https://science.sciencemag.org/content/359/6380/1094.summary *Science* 09 Mar 2018:Vol. 359, Issue 6380, pp. 1094-1096 DOI: 10.1126/science.aao2998.