IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FAKE NEWS DETECTION USING MACHINE LEARNING

Associate Prof. DR. Mahendra Sharma*1, Assistant Prof. Mrs. Laveena Sehgal*2, Blaghul Rizwan*3, Md Zamin Zafar*4

*1 Associate Professor, Department of Information

Technology, IIMT College Of Engineering, Greater Noida, Uttar Pradesh India.

*2 Assistant Professor, Department of Information

Technology, IIMT College Of Engineering, Greater Noida, Uttar Pradesh India.

*3 Student, Department of Information

Technology, IIMT College Of Engineering, Greater Noida, Uttar Pradesh India.

*4 Student, Department of Information

Technology, IIMT College Of Engineering, Greater Noida, Uttar Pradesh India.

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 and 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 is 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 result.

INTODUCTION

Fake news contains misleading information that could be checked. This maintains lies about a certain startup in a country or exaggerated cost of certain services for a country, which may arise unrest for some countries like an Arabic spring. there are organizations like the House of Commons and the crosscheck project, trying to deal with issue as confirming authors and accountable. However, their scope is so limited because they depend on human mutual 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 the classifies fake articles and can be used and integrated with any system for future use.

RELATED WORK

1. Social Media and Fake News

Social media includes websites and programs that are devoted to forums, social websites, microblogging, social bookmarking and wikis. On the other side some researchers consider the fake news as a result of accidental issue such as educational shock or underwriting actions like what happened in Nepal Earthquake case. 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 includes lots of misinformation.

2. Natural language processing

The main reason for utilizing natural language processing is to consider one or more specializations of system or an algorithm. natural language processing (NLP) painting 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. There has been 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 taggers, chunking, and semantic role labeling made NLP good subject of search.

This sentiment analysis extracts emotions on a particular subject. Sentiment analysis is composed of extracting a specific term for a subject, extracting the sentiment and pairing the connection analysis. The sentiment analysis uses dual languages resources for analysis: Glossary of meaning and sentiment models database full for constructive and destructive words and attempts to give classifications on a level of –5 to 5. parts of speech taggers tool for languages such as European languages are being explored to produce parts of language taggers tools in different languages such as Sanskrit, Hindi 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. Parts of Sanskrit word "speak" specifically use the treebank

method. Arabic utilizes vector machine (SVM) uses a method to automatically identify symbols and parts of speech and automatically expose basic sentences in Arabic text.

3. Data mining

data mining techniques are categorized into two main methods which are: supervised and unsupervised. The supervised method utilizes the training information in order to foresee the hidden activities. Unsupervised data mining is an attempt 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.

4. Machine learning (ML) Classification

Machine learning (ML) is a class of algorithms that help software system achieve more accurate results without having to reprogram them directly data scientists categorize 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. There are six algorithms that are adopted in this paper for classifying the fake news.

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 at 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 are useful for data exploration and predict the target variable quite efficiently.

Decision Tree Pseudo-code

```
GenerateDecision Tree (Sample s, features F)
1. If stop_condition(S, F)= true then
   a. leaf = create_Node()
   b. Leaf.lable= classify(s)
   c. Return leaf
2. root= creat_Node()
3. root.testcondition= find_bestSplit(s,f)
4. v = {v l v apossible outcome of root.testconditions}
5. for each value v belong V
6. sv: = {s | root.testcondition(s) = v and s belong to S};
7. child = Tree_Growth(Sv,F);
8. Grow child as descent of root and lable the edge (root--->child) as v
```

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

6. **Random Forest**

Random forests 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 a large number of decision trees, 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.

The applicability of Random Forest is the 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 Psudo-code

To make n classifiers:

For I= 1 to n do

Sample the training data T random with replacement for Ti output

Build a Ti- containing root node, Ni

Call BuildTree (Ni)

end For

BuildTree (N);

If N includes instance of only one class, then returns

else

Select z% of the possible splitting characteristics at random in N

Select the features F with the highest information gain to split on

Create f child nodes of N, Ni …., Nf, where F has f possible values (F1, …., Ff)

For i = 1 to f do

Set the content of Ni to Ti, where Ti is all instances in N that match Fi

Call Buildtree(Ni)

end for

end if

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. Given a set of N features, SVM algorithm uses n dimensions 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 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_RFE_SVM (F[0...high], value, mid,high)

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. Simply put, Naïve 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 it's red color, swirls, and the diameter is closed 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, Naïve Bayes assumes that all these functions share a separate proof of the apples.

Naïve Bayes Equation

$$P(c \mid x) = \{P(x \mid c)\, P(c)\} \setminus P(x)$$

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

Where:

P (c | X) is the posterior Probability.

P (x | c) is the Likelihood.

P (c) is the Class Prior Probability.

P (x) is the Predictor Prior Probability.

Naïve 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;

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's size for random forest model is very large and it might result in overfitting. 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.

9. KNN (k- Nearest Neighbours)

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.

KNN Pseudo-Code

Classify (X, Y, x) // X: training data, Y: class label of X, x: unidentified sample

For I = 1 to m do

Calculate distance d (Xi, x)

End for

Calculate set (I) containing indices for k smallest distance d (Xi, x)

Return majority label for {Yi where I belongs I}

KNN falls in the category of supervised learning and its main applications are instrusion 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.

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.

11. Related Work on Fake News Detection

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 unproductive models that do not fit with the other current models.

Used naive Bayes classifiers to detect fake news by naive Bayes. This method was performed as a software framework and experimented with various records from Facebook, etc., resulting in an accuracy of 74%. The paper neglected punctuation errors, resulting in poor accuracy. Estimated various ML algorithms and made the researchers on the percentage of 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. Utilize be Naive Bayes Classifiers, discuss how to implement fake news discovery to different social media sites. They used Facebook, Twitter and other social media applications as a data source for news. Accuracy is very low because the information on This site is not 100% credible. Discuss misleading and discovering rumors in real time. It utilizes a novelty-based characteristic underived its data source from Kaggle. The accuracy average of this pattern is 74.5%. Clickbait and sources do not consider unreliable, resulting in lower resolution. Used to distinguish Twitter spam senders. Among the various models used are Naïve 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. identified fake news in different ways. The accuracy is limited to 76% as a language model. Greater accuracy can be achieved if a productive model is used. Aim to utilize machine learning methods to detect fake news three common methods are utilized through their research: Naïve Bayes, Neural Network and Support Vector Machine (SVM). Normalization technique is an essential stage in data cleansing prior machine learning is used to categorize the data. Two more advanced methods, the neural network and the machine vector (SVM) reached an accuracy of 99.90%.

In 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 suggested model is analyzed for accuracy and recall. The final results improved by up to 8% using a mixed false message detection model.

Examined how fake news was used in the 2012 Dutch elections on Twitter. She examines the execution of eight supervised machine learning classifiers in the Twitter data set. We assume that the decision tree algorithm works best for the data set used with an F score of 88%. 613,033 tweets were rated, of which 328,897 were considered genuine and 284136 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.

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 the linear SVM workbook. The highest accuracy is 92%.

3. Methodology

This section presents the methodology used for classification. Using this model, a tool is implemented for detecting the fake articles in this method supervised machine learning is used for classifying the data set the first step in this classification problem is data set collection phrase, followed by preprocessing, implementing feature selection, then perform the training and testing of data set and finally running the classifiers. Figure describes the proposed system methodology. The methodology is based on conducting various experiments on datasets using the algorithm described in the previous section named random forest, SVM and section 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.

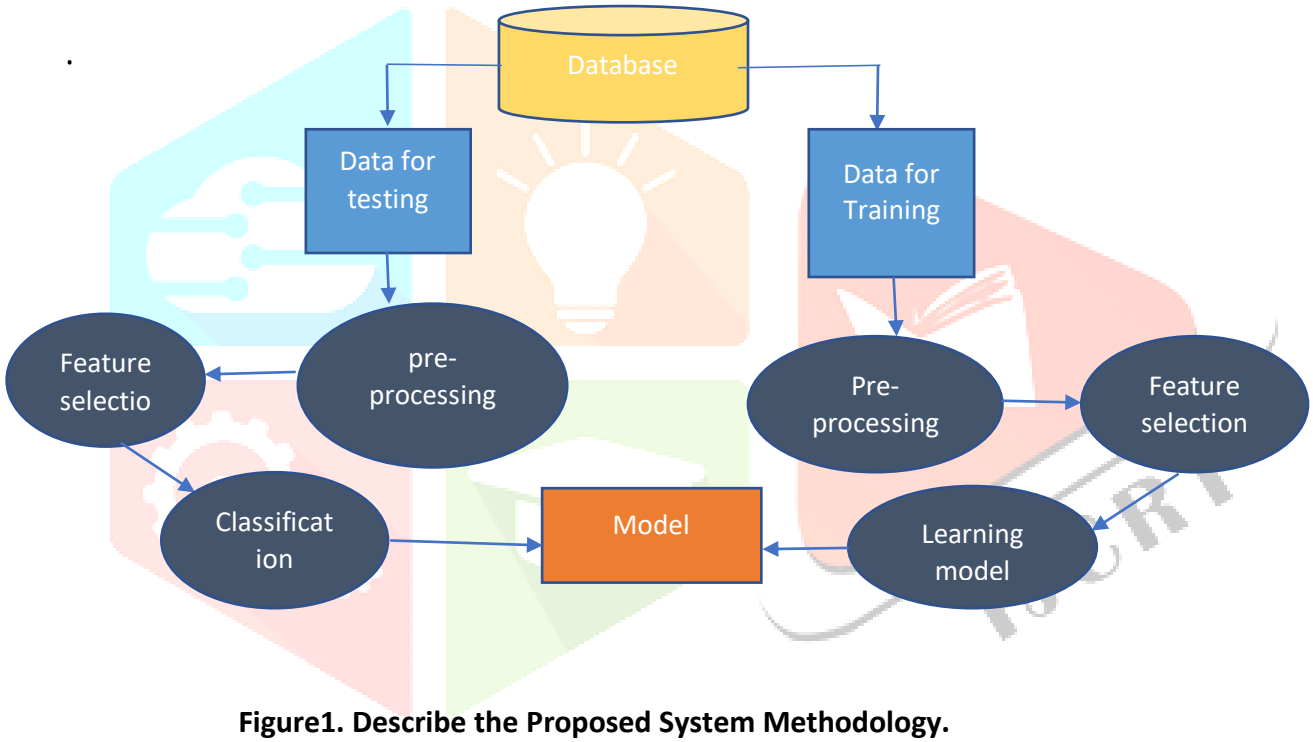

Figure1. Describe the Proposed System Methodology.

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 detections and embed the model in Python application to be used as a discovery for the fake news data. Also, appropriate refactorings have been performed on the Python code to produce an optimized code.

Massification algorithms applied in this model are K nearest neighbors, linear regression, XGBoot, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine. 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, the data set 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.

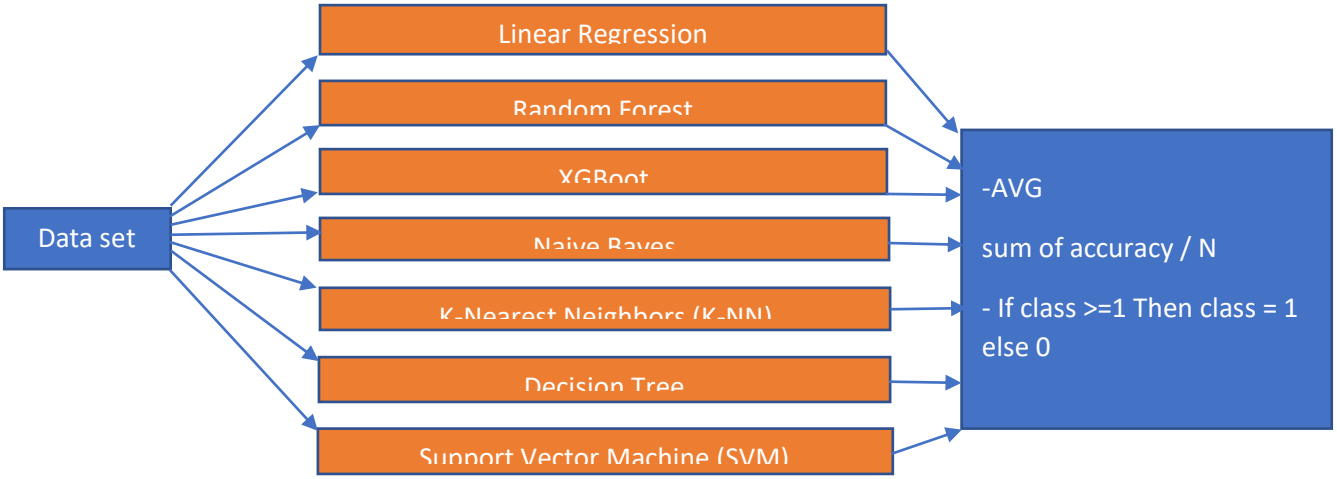

Figure 2. The Classification Algorithms

In the process of model creation, the approach of detecting political fake news is as follows: first step is collection political news data set (a liar data set 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 data set splitting apply ML algorithm. The figure 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 responds with N.B and Random forests are applied, then the model is created with response message. Testing is performed on the test data set, 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 the user. Full data set is created 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 data set to be used in our complete data set and leaving 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 we'll 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 the library contains useful tools like Count Vectorizer and Tiff Vectorizer. Data is viewed in graphical presentation with confusion matrix.

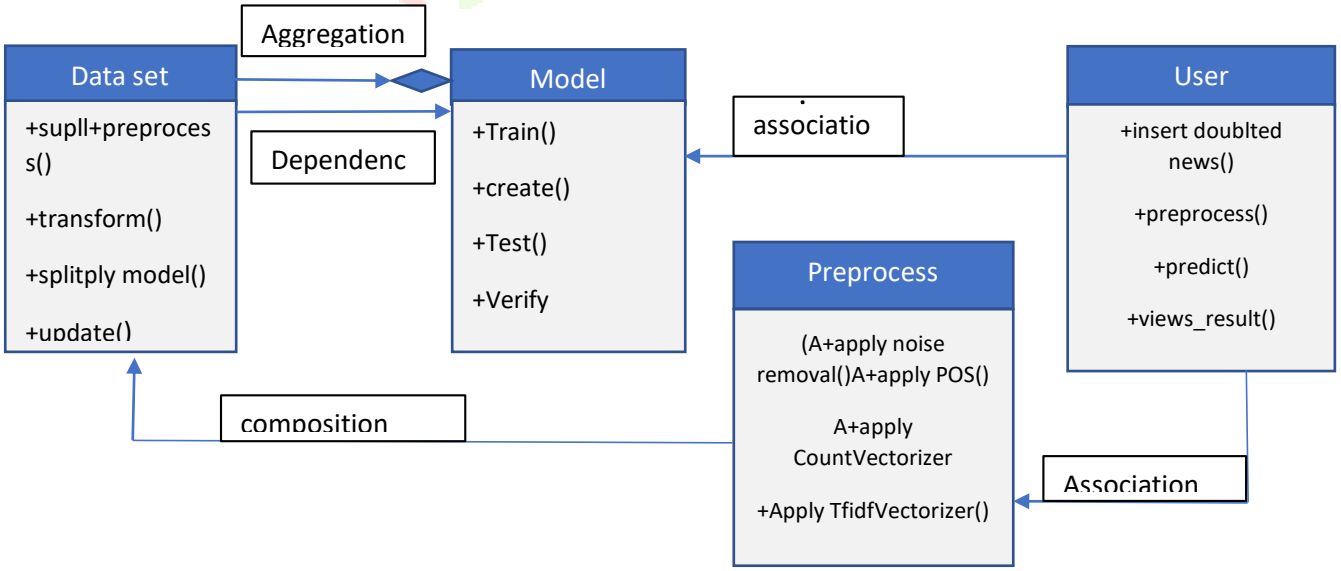

Figure 3. Fake Detector Model

RESULT

The scope of this project is to cover the political news data of a dataset known as Liar-dataset, it 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 6th algorithms used for the detections are as:

- XGBoot.
- Random Forest.
- Naïve Bayes.
- K-Nearest Neighbors (KNN).
- SVM.

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

CONCLUSION

The research in this paper focuses on detecting the fake news by reviewing it in two stages: characterizations and disclosure. in the first stage the basic concept and the principle 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 the displayed fake news detection approaches that is based on text analysis in the paper utilizes models based on speech characteristics and predictive model that do not fit with the other current models.

In the F4 mentioned research summary and system analysis we concluded that most of the research papers used aive based algorithm, and the prediction precision was between 70 to 76%, they mostly use qualitative analysis depending on sentiment analysis, titles, word frequency repetitions. In our approach we propose to add to these methodologies, another aspect, which is POS textual analysis, it is a quantitative approach its depends on adding numeric statical 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.

REFERENCES

1. https://www.researchgate.net/search?q=fake%20news%20detection
2. https://www.researchgate.net/search?q=fake%20news%20detection
3. https://www.researchgate.net/search?q=fake%20news%20detection
4. https://www.researchgate.net/search?q=fake%20news%20detection