International Journal of Civil, Mechanical and Energy Science, 6(2)
Mar-Apr, 2020 | Available: http://ijcmes.com/

Techniques of Fake News Detection

Harshit Garg[1], Ms. Alisha Goyal[2], Ms. Ankita Joshi[3]

[1]Student, Department of Computer Science & Engineering, Global Institute of Technology, Jaipur, Rajasthan, India
[2,3]Assistant Professor, Department of Computer Science &Engineering, Global Institute of Technology, Jaipur, Rajasthan, India

Abstract**—** *Fake news is the contents that claim people to believe with the falsification, sometime it is the sensitive messages. Mixing both believable and unbelievable information on social media has made the confusion of truth. That is the truth will be hardly classified. The techniques for detecting the Fake News means its a false story which comes from unauthorized source. Only by building a model based on a count vectorizer or a Term Frequency-Inverse Document Frequency i.e. TF-IDF score matrix calculation can only get you so far. It may happen that the meaning two article be similar. Combating the fake news is a classic text classification project with a straight forward proposition. We can implement a task by Naïve Bayes or any other method to find out the real vs fake news.*

Keywords**—Fake News, Machine Learning, Classifiers.**

I. INTRODUCTION

Different from the beginning of the internet, we produce more data and information than we are able to consume. Consequently, it is possible that some misinformation or rumours are generated and spread throughout the web, leading other users to believe and propagate them, in a chain of unintentional (or not) lies. Such misinformation can generate illusive thoughts and opinions, collective hysteria or other serious consequences. In order to avoid such things to happen, specially closed to political events such as elections, researchers have been studying the information flow and generation on social medias in the last years, focusing on subjects as opinion mining, users relationship, sentiment analysis, hatred spread etc. Based on a systematic review of recent literature published over the last 5 years, we synthesized different views dealing with fake news. We wanted to investigate machine learning applications to detect fake news, focusing on the characteristics of the different approaches and techniques, conceptual models for detecting fake news . In order to answer our questions and show the results of our work, we will present the definition of misinformation, hoax, fake news and its main common concept, meanwhile, systematically review a set of machine learning techniques used to detect such kind of information. We conclude outlining the challenges and research gaps in current state-of-art of automatic fake news detection.

II. APPROACHES TO DETECT FAKE NEWS

A. Artificial Intelligence :

The Rapid advances in technology have enabled medium to be published online and therefore the emergence of Facebook, Twitter, YouTube and other social networks. Social networks have become an important way for people to communicate with each other and share their ideas. The most important feature of social networks is that the rapid information sharing. In this context, the accuracy of the news or information published is extremely important. The spread of faux news in social networks has recently become one among the most important problems. Fake news affects people's lifestyle and social order and should cause some negativity. In this study, the foremost comprehensive and prestigious electronic databases are examined so as to seek out the newest articles about the detection of faux news in social networks by systematic literature review method. Due to use of artificial intelligence it checks approx 90% of fake news .so it's a important tool.

B. Suspicious News Detection Using Micro Blog Text:

We have new task that is use of micro blog text which supports to human for to detect suspicious fake article which is a crucial task along with costly. we have to find that article is fake or real for this task, we use various dataset and show results by using machine learning concept. It reduces human effort at a high rate.

C. Comparative Performance of Machine Learning Algorithm for Fake News Detection:

Fake news affect negativity to our society and public place which individual various problems and cause various mentality issues .The problem has been approached in this

paper from Natural Language Processing and Machine Learning perspectives. We can evaluate throught the dataset by use of a novel set of features extracted from the headlines and the contents. Performances of seven machine learning algorithms in terms of accuracies and F1 scores are compared. Gradient Boosting outperformed other classifiers with mean accuracy of 88% and F1-Score of 0.91.

III. DIFFERENT CLASSIFIERS TO DETECT FAKE NEWS

A. Random Forest Classifier :

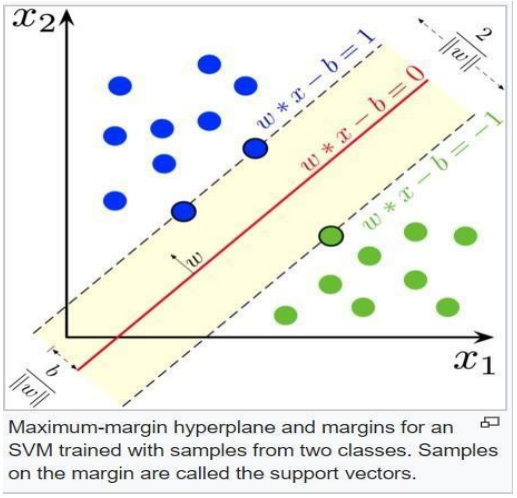
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a mess of decision trees at training time and outputting the category that's the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which, in Ho's formulation, may be a thanks to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark (as of 2019, owned by Minitab, Inc.). The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman so as to construct a set of decision trees with controlled variance.

B. Stochastic Gradient Descent:

Before talking about Stochastic Gradient Descent (SGD), let's first understand what's Gradient Descent? Gradient Descent is a very popular optimization technique in Machine Learning and Deep Learning and it can be used with most, if not all, of the learning algorithms. A gradient is essentially the slope of a function; the degree of change of a parameter with the quantity of change in another parameter. Mathematically, it can be described as the partial derivatives of a set of parameters with respect to its inputs. The more the gradient, the steeper the slope. Gradient Descent is a convex function. Gradient Descent are often described as an iterative method which is employed to seek out the values of the parameters of a function that minimizes the value function the maximum amount as possible. The parameters are initially defined a particular value and from that, Gradient Descent is run in an iterative fashion to find the optimal values of the parameters, using calculus, to find the minimum possible value of the given cost function.

C. Support Vector Machines Classifier :

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and multivariate analysis . Given a group of coaching examples, each marked as belonging to at least one or the opposite of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the opposite , making it a non- probabilistic binary linear classifier (although methods like Platt scaling exist to use SVM during a probabilistic classification setting). An SVM model may be a representation of the examples as points in space, mapped in order that the samples of the separate categories are divided by a transparent gap that is as wide as possible. New examples are then mapped into that very same space and predicted to belong to a category supported the side of the gap on which they fall.In addition to performing linear classification, SVMs can efficiently perform a non- linear classification using what's called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.When data are unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to seek out natural clustering of the info to groups, then map new data to those formed groups. The support-vector clustering algorithm, created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed within the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial application.



Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

Fig.1: SVM Analysis

a. K- Nearest Neighbor Classification:

K-NN is a supervised learning classification algorithm.

K-NN verifies similar things near to each other. In K-NN, K indicates number of nearest neighbors. Initially, select k value and group the data items into k groups based on similarity (distance). The items can be classified at the end. Distance can be calculated using Euclidean distance.

b. Decision Tree Classification:

One of the most widely used classifiers is Decision Tree Classifier. It is also a powerful classifier. Similar to SVM, Decision Tree can also perform both regression and classification. It is also a supervised learning algorithm. Decision Tree classifiers are more popular because tree analysis is easy to understand. It divides the given data set into small parts and a decision tree is incrementally constructed. The leaf nodes of a decision tree represent the classification. Decision trees are comfortable with numeric and categorical data.

IV. METHODOLOGY TO DETECT FAKE NEWS

A. Removal of Stop Words:

Remove stop words from the Parsed document that is from Text Document. After parsing the stop words will be removed from the sentences to produce high informative summary.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

B. Stemming of The Document:

Stemming is that the process of manufacturing morphological variants of a root/base word. Stemming programs are commonly mentioned as stemming algorithms or stemmers. A stemmer reduces the words “chocolates”, “chocolatey”, “choco” to the basis word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”.

Some more example of stemming for root "like" include:”.

->"likes"

->"liked"

->"likely"

->"liking"

C. Data Preprocessing:

In this phase, the dataset is taken as input from the kaggle. In the input dataset no missing value is there and the input dataset will be tokenized. The tokenized dataset will be processed and unwanted information will be removed from the dataset.

D. Classification:

In this phase, the input dataset will be divided into training and testing. The training dataset will be 60 percent of the whole data and 40 percent will be test dataset. The KNN is the k nearest neighbor algorithm which can calculate the nearest neighbor values in the input dataset data needs to be classified. The random forest highly efficient algorithm which can provide greats results even without the presence of hyper-parameter tuning is called random forest algorithm. Due to its high simplicity and the fact that both classification and regression tasks can use it this classifier is gaining huge popularity. In the training time, multitude of decision tree is calculated and the mean prediction of individual trees is given as output.

V. CONCLUSION AND FUTURE SCOPE

Truth discovery is the problem of detecting true facts from multiple conflicting sources. Truth discovery methods do not explore the fact claims directly, but rely on a collection of contradicting sources that record the properties of objects to determine the truth value. Truth discovery aims to determine the source credibility and object truthfulness at the same time. The fake news detection problem can enjoy various aspects of truth discovery approaches under different scenarios. First, the credibility of different news outlets can be modeled to infer the truthfulness of reported news. Second, relevant social media posts can also be modeled as social response sources to better determine the truthfulness of claims. However, there are another issues that has got to be considered to use truth discovery to fake news detection in social media scenarios. First, most existing truth discovery methods focus on handling structured input in the form of Subject-Predicate-Object with the increasing popularity of social media, more and more people consume news from social media instead of traditional news media. However, social media has also been wont to spread fake news, which has strong negative impacts on individual users and broader society. In this article, we explored the fake news problem by reviewing existing literature in two phases: characterization and detection. In the characterization phase, we introduced the basic concepts and principles of fake news in both traditional media and social media. In the detection phase, we reviewed existing fake news detection approaches from a knowledge mining perspective, including feature extraction and model construction. We also further discussed the datasets, evaluation metrics, and promising

future directions in fake news detection research and expand the field to other applications.

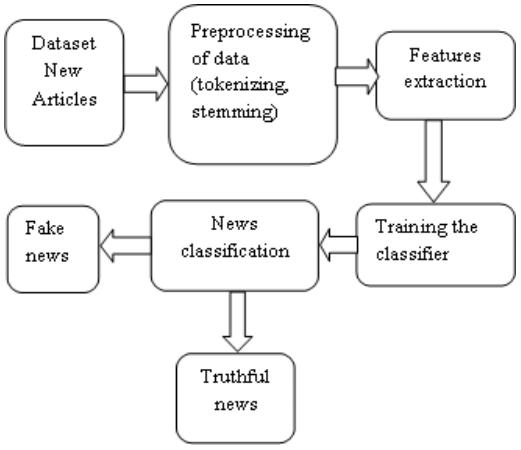


Fig.2: flow chart for detecting fake news

The nearest neighbor values of the dataset are calculated using the Euclidian distance formula. The number k value is selected from the network and based on k number of values the data can be classified into certain classes. The number of hyper planes depends upon the number of classes.

REFERENCES

[1] Nir Kshetri, Jeffrey Voas, "The Economics of Fake News", IEEE, IT Professional, 2017, Volume: 19, Issue: 6, Pages: 8 - 12

[2] Roger Musson, "Views: The frost report: fake news is nothing new", 2017, IEEE, Astronomy & Geophysics, Volume: 58, Issue: 3, Pages: 3.10 - 3.10

[3] Hal Berghel, "Oh, What a Tangled Web: Russian Hacking, Fake News, and the 2016 US Presidential Election", IEEE, Computer, 2017, Volume: 50, Issue: 9, Pages: 87 - 91

[4] Hal Berghel, "Alt-News and Post-Truths in the "Fake News" Era", IEEE, Computer, 2017, Volume: 50, Issue: 4.