Fake News Detection with Machine Learning

Jayesh Patel, Melroy Barreto, UtpalSahakari, Supriya Patil

Abstract*: As the internet is becoming part of our daily routine there is sudden growth and popularity of online news reading. This news can become a major issue to the public and government bodies (especially politically) if its fake hence authentication is necessary. It is essential to flag the fake news before it goes viral and misleads the society. In this paper, various Natural Language Processing techniques along with the number of classifiers are used to identify news content for its credibility.Further this technique can be used for various applications like plagiarismcheck , checking for criminal records.

Keywords —K-Means Cluster (K-means), K-Nearest neighbor (KNN), Stochastic Gradient Descent (SGD),Support Vector Machines (SVM).

I. INTRODUCTION

Objective— To detect the fake news from the circulated abstract, or a composite document with contradicting, misleading statements. It's necessary to build a neural network (model) that can differentiate information between True aspects - "Legit news" and False aspects -"Fake news". The trend of fake news has increased drastically in this digitally connected world causing widespread confusion among people as to what is real and fake. Fake news spreads like wildfire impacting millions of people in everyday life. The fake news is sometimes deliberately spread so as to spread hatred, promote illegal deeds or spoil some events. Unauthenticated false news can also produce such havoc in public masses. Fake news detection on new fast happening events is one of the major setbacks and challenges on social media platforms. Tomas Mikolov [3] has used various ways of representing a dataset as word vectors. The word2vec representation is obtained by using "word embeddings", where each word in a news content is converted into a dense vector of floating-point values (the length of the vector is a parameter you specify). This method (word embeddings) was adopted to overcome the shortcomings of some earlier used methods like the "one-hot encodings" or representing each word as a number, but even though these methods prove to be reliable do not guarantee efficiency since each word is represented as a vector and for a document consisting of millions of words will have millions of vector representations which will consume lot of time and also memory. Xin Rong [4] has elaborated more on the "skip-gram and continuous bag of words methods of data representation"which inevitably face problems of memory and time consumption.

Revised Manuscript Received on November 01, 2020.

Jayesh Patel, Associate professor in Electronics and Telecommunication Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. India

MelroyBarreto, Associate professor in Electronics and Telecommunication Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. India

UtpalSahakari, Associate professor in Electronics and Telecommunication Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. India

Dr. Supriya Patil, Associate professor in Electronics and Telecommunication Engineering Department at Padre Conceicao College of Engineering, Verna, Goa. India

To overcome this disadvantage Doc2Vec model can be used which is just an addition of a "paragraph id" to the word2vec representation wherein sentences or paragraphs can be represented as vectors. Real time news cannot be distinguished or perhaps the trusted source itself might give fake news at most periods of time. Hence it is necessary to monitor the sources for a particular period and also at the same time broaden the coverage of sources. In this paper, the identification of fake news is accomplished by using various Natural Language Processing Techniques along with the classifiers such as **Support Vector Machines (SVM), K-Means Cluster (K-means), K-Nearest neighbor (KNN) and Stochastic Gradient Descent (SGD).** The paper will be organized as follows: data set details in part II, followed by pre-processing techniques in part III, feature extraction method using Doc2Vec model in part IV, various ANN training algorithms in part V, results in part VI and conclusions.

II. DATASET

This paper utilizes the Kaggle database for the detection of fake news. Data set has 10349 fake and 10369 non-fake news respectively. About 16,000 samples are used for training and 4000 for testing. The various attributes of the database for every news are identification number, title, author, text and label associated with it

III. DATA PRE-PROCESSING

For efficient processing of the data using Natural Language Processing technique, pre-processing of the data is implemented before feature extraction. Feature extraction using 'word embedding' technique, the text is converted in the form that the machine can understand. But conversion of thousands of pages of text content can be very time consuming and decreases performance memory wise hence before applying Doc2Vec technique the news document is processed to remove stopwords (conjunction, prepositions, articles, etc.), delete special characters and finally the text in the article is converted to lowercase. It produces a comma-separated list of words, which are further processed using Doc2Vec method of feature extraction.

IV. FEATURE EXTRACTION

The objective of Doc2Vec is to create a numeric representation of a given document, irrespective of its length. But unlike words, document materials do not come in reasoning or logical structures such as words, so another careful way has to be discovered. Doc2Vec is a model developed in 2014 which is heavily based on the earlier Word2Vec model, which creates vector representations (of floating point values) for words.

Word2Vec represents documents by combining the vectors of all the individual words, but in this process, it loses all the word order information. Doc2Vec is an extension of Word2Vec, adding a 'paragraph id' (***D***) to the output representation, which contains paragraph specific information about the document as a whole, and allows the ***nn***model to learn extra information about word ordering. Doc2Vec preserves word-order data which makes it feasible for fake news detection, since the goal is to detect precise differences between text articles.

The doc2vec may be used in the following way: for training, a set of articles are required. A word vector ***w*** is generated for each word, and a document vector ***d*** is generated for each article/document. This model trains automatic weights for a softmax hidden layer where activation function is formulated. In the inference stage, all weights are initially fixed to calculate the document vector hence a new document may be presented.

V. TRAINING MODELS

A. SVM

'Support vector machine' is a set of supervised learning methodsusedfor classification ,regression and outlier detection. To separate the data points into two classes, there are possible planes (hyperplanes) that could be chosen. Consider example classification of dog and cat .The purpose is to separate them both based on characteristic data points like shape of ears, tail ,paws which in terms of algorithm a plane of separation the distance between characteristic data points of both classes is termed as margin and the maximum it is the better classification. Refer Fig.1 Also, the number of features of classification controls the hyperplane. The hyperplane is just a line or linear if the number of features is 2. The hyperplane becomes a two/multi-dimensional plane if features of classification or data points are two or more respectively.

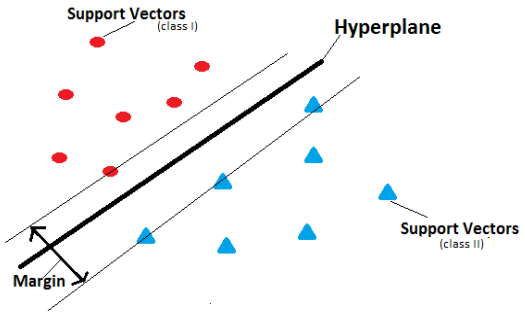

Fig1 -SVM

Maximizing the margin distance between them can enable us to classify the input data with more confidence.

Kernel is used for data point's separation and to determine the hyperplane. Also, sometimes gamma function is used in formulas to simplify the mathematical calculation and more corrective classification. Following are some of the functions used as per characteristic data input points.

LINEAR KERNEL

The formula of linear kernel is as below usually dot product between any two pairs of observed input values x, xi.−

$$K(x,xi)=\sum(x * xi)$$

POLYNOMIAL KERNEL

This form of linear kernel type helps us differentiate curved or nonlinear type space with input x, xi respective inputs. Following is the formula for polynomial kernel −

$$K(x,xi)=1+\sum(x * xi)\text{^}dK(x,xi)=1+\sum(x * xi)\text{^}d$$

Here D is the degree of polynomial, which we need to specify manually in the learning algorithm.

RADIAL BASIS FUNCTION (RBF) KERNEL

RBF kernel is mostly used in SVM classification, when maps input space is indefinite dimensional space. Following formula explains it mathematically −

$$K(x,xi)=\exp(-\text{gamma} * \sum(x-xi^2)K(x,xi))$$
$$=\exp(-\text{gamma} *\sum(x-xi^2))$$

Here, gamma value ranges 0 - 1.A default value of gamma is 0.1 if not stated manually.

B. K-means

K-means clustering is one of the most commonly used "unsupervised machine learning algorithms for partitioning a given data set into a set of k groups (i.e. *k clusters*), where k represents the number of groups pre-specified by the analyst". It classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e., high *intra-class similarity*), whereas objects from different clusters are as dissimilar as possible (i.e., low *inter-class similarity*). In k-means clustering, each cluster is represented by its center (i.e., *centroid*) which corresponds to the mean of points assigned to the cluster.

The procedure followed is simple as followed up next. It's an easy way to classify a given data set into a certain number of clusters (assume k clusters). The main aim is to define k centers, for these clusters. These k-centers should be placed in a definite way because different locations produce different results. The farther these k-centers are placed the better results can be expected. The next step is to take each point belonging to a given data set and associate it to the nearest center. No free points indicate completion of the first step and an early group classification is achieved by this way. Now we need to re-calculate k new centroids of the clusters resulting from the previous steps. After we have these k new centroids, a new connection link has to be set up between the same data set points and the nearest new center. A loop has been formulated. As a result of this loop the k centers change their location step by step eventually until no more changes are done or in other words centers become fixed and don't move. K-means algorithm's objective is minimizing an objective function know as squared error function given by the formula

$$J(L)= \sum_{i=1}^{n}\sum_{j=1}^{ni}\left(\left|\left|Ki-Lj\right|\right|\right)^2$$

Where, '||Ki - Lj||' is the Euclidean distance between Ki and Lj.

'ni' is the number of data points in the cluster.

'n' is the number of cluster centers.

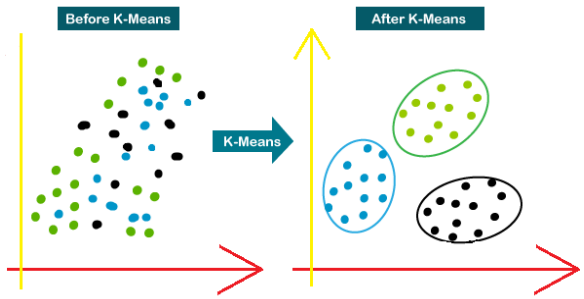

Fig 2 –K-Means

Algorithmic steps for k-means clustering:
Let K = {k1,k2,k3,……..,kn} be the set of data points and L = {l1,l2,…….,ln} be the set of centers.
1. Randomly select 'n' cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is the minimum of all the cluster centers.
4. Recalculate the new cluster center using:
li=(1/ni)j=1niKi
Where, 'ni' represents the number of data points in the cluster.
5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3.

C. KNN

KNN is a supervised learning algorithm that can be used to solve classification and regression problems. It is a non-parametric which uses a database of many classified points to predict the classification of new sample point. It is termed as a lazy algorithm as it does not do any generalization or training for data points. This suggests that the training process is pretty much fast. Lack of generalization means that the KNN training phase is minimal or it keeps all the training data. The **k** value should always be chosen as an odd value.

Algorithm steps for knn:
1. Input the sample point to be classified.
2. Assume value for 'k' nearest neighbors as required.
3. Find the Euclidean distances between all the elements of classification.
4. Depending on the value of 'k' find the k elements with the least Euclidean distances.
5. Apply simple majority.
6. Plot the sample point to the majority nearest neighbor classification.

To find the distances between the sample point and the labeled points of the database a metric has to be defined. The most common metric used is Euclidean. Some of the others include Euclidean squared, City-block, and Chebyshev:

D(x,p)={$\sqrt{(x-p)2}$Euclidean
$(x-p)^2$ Euclidean squared
Abs(x-p) Cityblock
Max (|x-p|) Chebyshev}

After selecting the value of k, you can make predictions based on the KNN examples. For regression type, KNN predictions is the average of the k-nearest neighbor's outcome.

$$Y=\sum_{i=1}^{k} yi$$

Where yiis the ith case of the examples sample and Y is the prediction (outcome) of the query point and k is the total number of samples.

D. Stochastic Gradient Descent (SGD)

In SGD, the output unit compares actual output to determine the error associated with that pattern. Based on the error delta K (∂_k) factor is calculated and it is used to distribute error at the input layer back to all units in the input layer. The Multiple neural networks used for SGD algorithm is as shown in the FIG.3. This algorithm consists of 2 basic steps.

- W_{kj}- represents weight from the hidden to the output layer.
- W_{ji}- are weight from the input to the hidden layer.
  - a- 'an activation value'.
  - T_k- represents a target value.
  - Net- the net input.

1. Determine connection weights in output layers.

$\Delta W_{kj} = \varepsilon\delta_k a_j$;
$\delta_k = (T_k - a_k)*a_k*(1 - a_k)$

2. Determine connection weights in hidden layers.

$\Delta w_{ji} = \varepsilon\delta_j a_i$:
$\Delta j = \sum(\delta_k W_{kj})*a_i*(1 - a_i)$

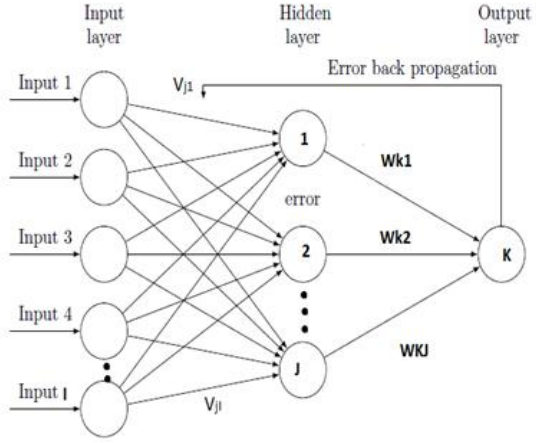

Fig 3 - SGD

Algorithm working

1. Calculated the feed-forward signals from the input to the output.
2. Calculate output error based on the predictions and the target.
3. Back propagate the error signals by weighing it by the weights in previous layers and the gradients of the associated activation functions.
4. Calculating the gradients for the parameters based on the back propagated error signal and the feedforward signals from the inputs.
5. Update the parameters using the calculated gradients.
6. SGD works continuously based on mathematically derived formulas. So, some criteria or condition has to be found to stop this continuous loop. Local minima are intended for this purpose. At each loop the errors are scaled based on its gradient weight.

The error with minimum gradient weight referred to as local minima is chosen and included in readings. We can't avoid errors because it will raise anyway. So, this algorithm's idea is to choose minimum gradient weight error so that we can improve efficiency of the algorithm.
Overall Goal of the algorithm is to modify the weight in the network such that the output vector is as close as possibleto the desired output vector.

VI. RESULTS

The comparison of classification accuracy obtained by implementation of preprocessing methods and classification algorithms such as SVM, K-Means, KNN and SGD is as shown in Table 1.

Table 1 Simulated Algorithms and their Accuracy

ALGORITHM	ACCURACY (learning rate)
SVM	88.47%
K-MEANS	40.37%
KNN	86.90%
SGD	90.32% (constant) 90.42% (adaptive)

VII. CONCLUSION

Fake news detection is a great help when it comes to various applications such as plagiarism check, verification of criminal records and also as a security option for various social media applications. Stochastic gradient descent proved to be the best algorithm to predict fake news correctly with an adaptive type of learning behavior with 90.32% for constant type of learning rate and 90.42% for adaptive type of learning. SVM and KNN also gave good results, if not best, with an accuracy of 88.47% and 86.90% respectively. While K-means was the worst and did not serve any purpose with a very low accuracy of fake news prediction.

REFERENCES

1. Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, David Lazer, RESEARCH ARTICLE Fake news on Twitter during the 2016 U.S. presidential election, Science 25 Jan 2019:Vol. 363, Issue 6425, pp. 374-378 DOI: 10.1126/science.aau2706 Fake News.
2. Dataset reference: Kaggle https://www.kaggle.com/c/fake-news/data
3. Distributed Representations of Words and Phrases and their Compositionality by Tomas Mikolov.
4. *word2vec Parameter Learning Explained* by Xin Rong https://arxiv.org/pdf/1411.2738 .
5. Hieu Pham and Daniel Boley, An Empirical Approach to Sentiment Analysis with Doc2Vec, University of Minnesota, Computer Science and Engineering, doc2vec:https://github.com/asprenger/doc2vec.
6. Doc2vecdocumentation:https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html#sphx-glr-auto-examples-tutorials-run-doc2vec-lee-py.
7. Theodoros Evgeniou and Massimiliano Pontil,WORKSHOP ON SUPPORT VECTOR MACHINES: THEORY AND APPLICATIONS,Center for Biological And Computational Learning, and Artificial Intelligence Laboratory,MIT, E25-201,Cambridge, MA 02139,USA.
8. DURGESH K. SRIVASTAVA,LEKHA BHAMBHU,DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE, Journal of Theoretical and Applied Information Technology.
9. Oyelade, O. J, Oladipupo, O. O,Obagbuwa, I. C, Application of k-Means Clustering algorithm for prediction of Students' Academic Performance, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010.
10. Imad Dabbura (Sep 17,2018) K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks, Towards datascience_ref:https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a.
11. Yun-lei Cai, Duo Ji, Dong-feng Cai, A KNN Research Paper Classification Method Based on Shared Nearest Neighbor, Proceedings of NTCIR-8 Workshop Meeting, June 15–18, 2010, Tokyo, Japan.
12. Sebastian Ruder, An overview of gradient descent optimization algorithms, arXiv:1609.04747v2 [cs.LG] 15 Jun 2017.
13. Fernando Cardoso Durier da Silva, Rafael Vieira da Costa Alves, Ana Cristina Bicharra Garcia, Can Machines Learn to Detect Fake News? A Survey Focused on Social Media, Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019URI:https://hdl.handle.net/10125/59713ISBN: 978-0-9981331-2-6 (CC BY-NC-ND 4.0) Page 2763.

AUTHORS PROFILE

Jayesh Patel, completed Bachelor of Engineering in Electronics & Telecom Engineering from Padre Conceicao College of Engineering, Verna-Goa.Currently working as Junior Software Engineer.

MelroyBarreto, completed B.E in Electronics and Telecommunications Engineering from Padre Conceicao College of Engineering, Verna-Goa. Currently working in the field of web development.

UtpalSahakari, completed Bachelor of Engineering in Electronics & Telecom Engineering from Padre Conceicao College of Engineering, Verna-Goa.Currentlyworking in the field pcb design and manufacturing. .

Dr. Supriya Patilis an Associate professor in Electronics and Telecommunication Engineering Department at Padre Conceicao College of Engineering, Verna,Goa. She received B.E (Instrumentation), M.E. (Electronics) degree from Shivaji University and Ph.D (Electronics) from Goa University. She worked for her Ph. D in the area of microarray-based cancer classification. She has 24 years of teaching experience with specialization in Signal Processing and Artificial Neural Network. She has authored and co-authored over twenty conference/journal papers in field of Artificial Neural Network.