https://doi.org/10.31871/IJNTR.6.10.16

Performance of Machine Learning Algorithm for Forged News Recognition

Kumar Divyanshu, Dr. M S Rudramurthy

Abstract— **News consumption from online network is very dangerous now days. From one point of view, its minimal effort, easy access, and rapid spread of data lead individuals to seek out and devour news from web based life. Then yet again, it empowers the extensive spread of Forged news, i.e. low quality news with persistently phony data. The purpose of this research paper is to detect the forged news by various machine learning algorithm (Naïve Bayes Classifier, Support vector machine, Feed Forward neural network) and compare the accuracy of the learning algorithm on two different system configurations with same libraries and datasets.**

Index Terms— **Machine learning; Forged news; Feed Forward neural network; Support Vector Machine; Naïve Baye Classifier.**

I. Introduction

Forged news is substantially hot debatable socio-political issues in recent couples of years. Legitimacy of information is dilemma queue and engaging the major remarkable population in nano second. Forged news or hoax news is disinformation spread through the media especially social media. This counterfeit news is gradually spinning into a danger to our general public. It is regularly produced for big business benefit to pull in watchers and collect promoting profits. Be that as it may be individuals or organizations with potentially hateful agenda have been known to instigate forged news in order to influence events and arrangements around the globe in their favour. That's why it is essential to promote studies in order to prevent and tackle false news so that they cannot be considered a threat to society. There are no instant breakers for Forged news detection, but it's the most needed thing to be added in the digital content management system. Need an idealistic technical solution to do the same and machine learning models have a past predictive success record to detect the originality of news. Various high powered predictive models such as Naïve Bayes, support vector method, feed forward neural network are used for predicting whether information content is Forged or real.

II. Related Work

A. Feed Forward Neural Network

Yaov Goldberg [2] proposed that within a previous couple of decades neural network emerged as powerful machine learning models producing innovative effects in areas like image recognition and language processing. Now neural system began to be implemented to textual natural language that shows potential success. The discussion covers various neural network which was quite vital as it helped in picking different algorithm to use in my project.

B. Naïve Bayes Classifier and Support Vector Machine

Z. H. Moe et.al [10] proposed the comparison performance of NBC and SVM on Document classification. The system they developed calculates the accuracy of testing data using holdout method. This helps me to include NBC and SVM for forged news recognition.

C. Detecting and Preventing Clickbaits in Online News Media

Abhijnan Chakraborty et.al [6] proposed that most news websites generates revenue from the readers visiting their websites and due to numerous website they are competing each other for reader's attention by generating catchy headlines. They build a browser extension to warn the user with misleading headlines.

D. Effects of spreading Forged news

Hunt Allcott et.al [4] proposed that how forged news can affect the large population in making decision by taking example of U.S. presidential election.

Meital Balmas [5] explains that what if fake news becomes real and its political effects. The discussion covers manipulation during 2006 Israel election campaign and also demonstrated that perceived realism of fake news is stronger among individuals.

III. Machine learning algorithm and implementation

Supervised Machine learning is implemented which comprises of label data divided into training data set and testing data set.

A. Naïve Bayes Classifier

NBC is an easy but unexpectedly strongest algorithm for predictive modeling. It relies on Bayes theorem which is conditional probability by this we will discover the chances of an event could occur given the data of the past event. We utilized the scikit-learn execution of Gaussian Naïve Bayes in which a probabilistic approach is used with the hypothesis. Select the hypothesis with the greatest probability which is called as maximum probable hypothesis. NVC are frequently used in sentiment analysis, spam filtering and direction systems. They are instant and easy to implement but their biggest con is the need for predictors to be free from bias.

Kumar Divyanshu, PG Scholar, Department of Information Science & Engineering, Siddaganga Institute of Technology, Tumakuru, India.

Dr. M S Rudramurthy, Associate Professor, Department of Information Science & Engineering, Siddaganga Institute of Technology, Tumakuru, India.

B. *Support Vector Machine*

SVM is a supervised machine learning algorithm which may be useful for both classification and regression purpose. We utilize the Radial Basis function in our project. The most crucial aim of a support vector machine would be to segregate the offered data from the most elegant way possible. After the segregation is completed, the length between the nearest points is referred to as the perimeter. The method is to pick a hyperplane with the most potential margin between your service vectors from the presented data collections. To divide the two types of points, you'll find many potential hyperplanes that can be chosen. Our purpose is to seek out a plane with the most perimeters, i.e., the most space between data points of the groups. Assessing the perimeter space stipulates some reinforcement; therefore, prospective data points might be categorized with greater optimism.

C. *Feed Forward Neural Network*

These networks are called feed-forward because the information only move forwards in the neural network, through the input nodes afterwards throughout the hidden layers (single or many layers) and finally throughout the nodes. The hidden layers are between the Output and Input layers, so since the practice, data will not demonstrate exactly the desired output for all these layers. A network may comprise any amount of hidden layers with any range of hidden units. A unit ostensibly looks like a neuron that takes input in components of prior layers and also simplifies its detection value. Neurons in each layer execute the similar function as human brain neurons do. Neural networks approximate the arrangement of your brain. A neural system structure has been coordinated to layers. In contrast, each coating includes lots of simple processing components, nodes farther attached to several nodes from the layers below and above. The data will be fed in the bottom layer that's subsequently relayed into the next coating. Unlike humans, artificial neural networks have been fed up with a massive number of data to the master. While artificial neural rhythms were initially built to be the neural networks but neural activity inside our brains is a lot more technical than could be indicated simply by analyzing artificial mice. Neuroscientists suggest that neurons tend not to arrive in an outcome signal by summing up the inputs. Additionally, real neurons tend not to remain before inputs shift, and also the presses can encode information with complex heartbeat structures. Brain-inspired metaphor while its name implies, neural-networks are motivated by the mind computation mechanism, and that contains computation units called nerves. From the event, a neuron can be just a computational unit that's scalar inputs and inputs. The nerves are associated with one another, forming a system: the outcome of a neuron can feed in to the inputs of a couple of neurons. Such networks were demonstrated to be somewhat competent computational apparatus. When the weights have been placed properly, a neural system with sufficient nerves and also a non linear activation function can approximate an extremely wide assortment of mathematical purposes. We executed a feed-forward neural system model utilizing Tensor flow. Neural networks are normally used in present-day NLP applications [2], rather than more established methodologies which centered around straight models, for example, SVM's and strategic relapse. Our neural system executions utilize three shrouded layers. In the Tensor flow execution, all layers had 300 neurons each, mixed with dropout layers to abstain from over-fitting. For our actuation work, we picked the Rectified Linear Unit (ReLU), which has been found to achieve well in NLP applications. The model we designed can be summarized using the following layers: Input Layer, dense1 Layer, dropout1 Layer, dense2 Layer, dropout2 Layer, dense3 Layer, dropout3 Layer, logits Layer. Here, The input layer is responsible for receiving vectors from Doc2Vec, and the Dense layer implements the operation: output equals to activation*dot (input, kernel) + bias, and dropout layers are used to avoid over fitting. In the dropout phase, some neurons are dropped during training. Logit is used to map the probability of the words occurring to the validity of the article.

D. *Proposed Scheme*

Pre-Processing: The embeddings utilized for the majority of our screens are produced using the Doc2Vec version. The purpose is to make a vector representation of each report. Before applying Doc2Vec, we do several simple pre processing of this data. Including removing stop words, deleting special personalities and accentuation, and shifting entire content. This produces a comma-isolated run down of provisions, which is a donation into this Doc2Vec calculation to deliver a 300-length Shifting vector for each report.
Training: All the models are programmed using python with anaconda support. Each algorithm is treated as a separate module and is trained isolated. The input data is passed into the algorithm as a vector format. This data is then analyzed to provide appropriate weights to the training algorithm. Optimization algorithms are used to avoid under fitting and over fitting. When the model attains a threshold loss function attribute the training is terminated, and the model is considered fit.

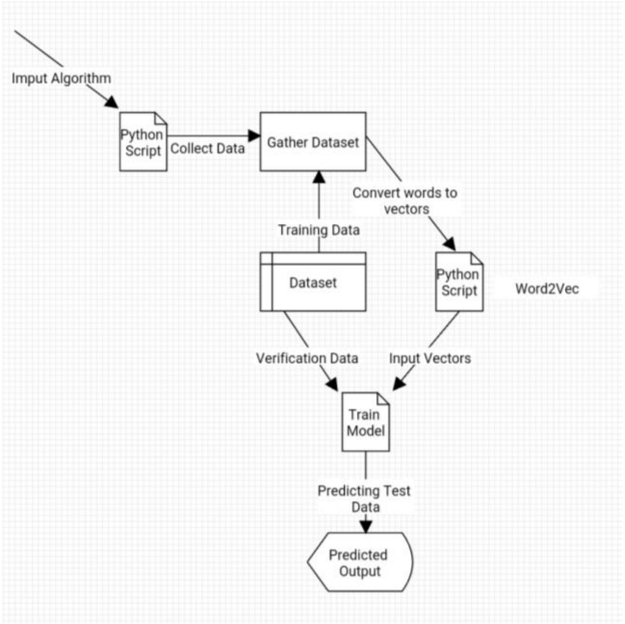

Fig (1): Flow Diagram

Predictions: Now, the trained model is used to predict the results of air-gapped data. The performance of all the algorithms of this data is stored in local memory for comparison.

Comparison: Performance Matrix of the data is visualized in a graphical format after execution on both the systems i.e. system one (i3 @2GHz, 4 GB ram) and system two (AMD A8 with R5 graphics @2.2GHz, 8 GB ram).

IV. Results

A. *Naïve Bayes Classifier*

On both systems NBC gives the same accuracy which is 72.16% this shows that NBC is independent on hardware. In Fig 2, confusion matrix used to describe the performance of a NBC on test datasets.

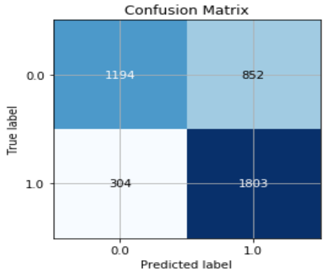

Fig (2): Confusion matrix of NBC (s1, s2)

B. *Support vector Machine*

In case of SVM accuracy varies as system configuration changes from s1=91.76% to s2=88.15%. In Fig 3a,3b it shows the performance of SVM on test datasets.

- System 1

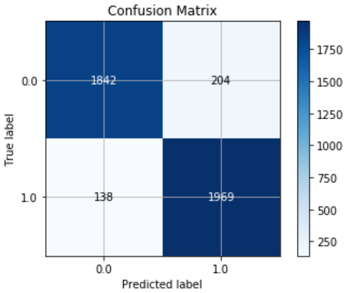

Fig (3a): Confusion matrix of SVM s1

- System 2

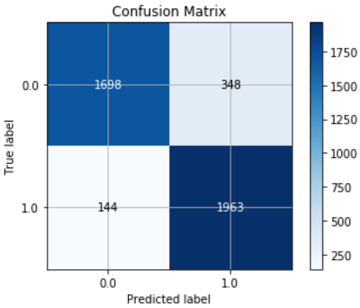

Fig (3b): Confusion matrix of SVM s2

C. *Feed Forward Neural Network*

In case of FFNN it totally depends on number of dataset i.e. if the model is trained on 26o thousand datasets then it gives more accurate result on test datasets which is 89.72% as shown in Fig 5a and Fig 5b. Rather the model trained on 80 thousand datasets which gives the accuracy of 86.30% as shown in Fig 4a and Fig 4b.

- System 1

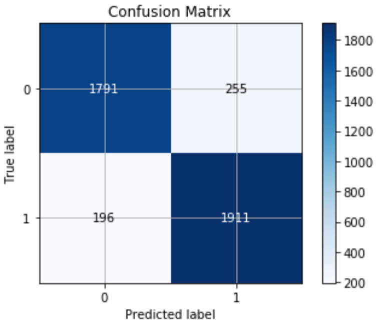

Fig (4a) : Confusion matrix of FFNN s1

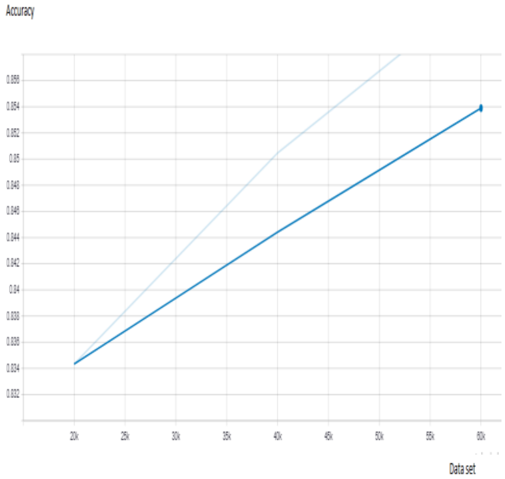

Fig (4b) : Accuracy of FFNN s1

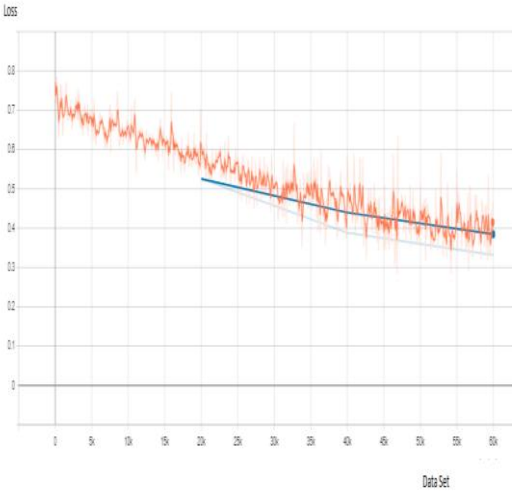

Fig (4c) : Loss of FFNN s1

- System 2

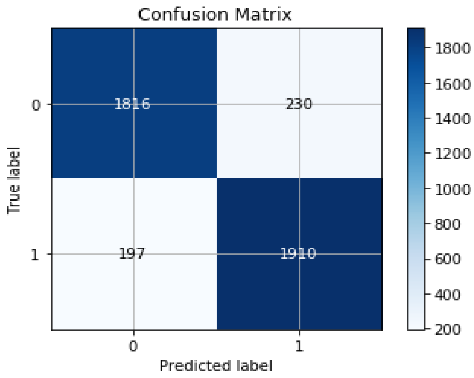

Fig (5a) : Confusion matrix of FFNN s2

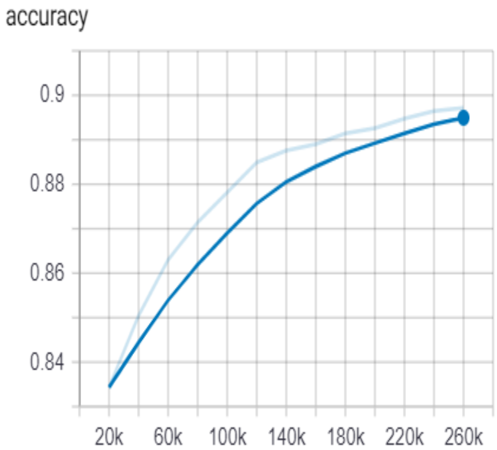

Fig (5b) : Accuracy of FFNN s2

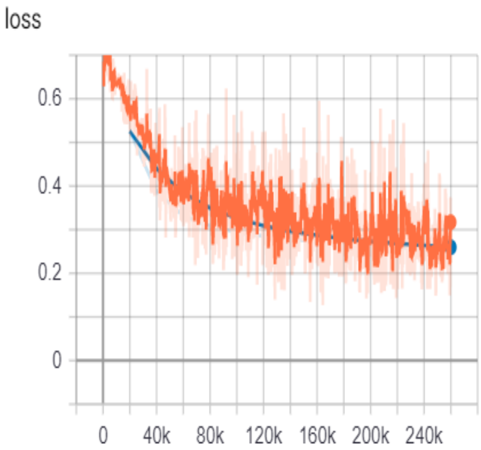

Fig (5c) : Loss of FFNN s2

As shown in Table 1, NVC accuracy is least and same on different systems but SVM accuracy decreases if system configuration is increased and FFNN accuracy increases by increasing the train datasets steps processed and it does not depend on system configuration but depend on number of time it is trained on train datasets to predict on test datasets.

Table 1 Shows the final comparison

Sl No.	**Accuracy**		
	ML Algorithms	***System 1***	***System 2***
1.	Naïve Bayes Classifier	72.16%	72.16%
2.	Support vector machine	91.76%	88.15%
3.	Feed Forward neural network	86.30%	89.72%

V. Conclusion

Different machine learning methods used in entire detection and prediction method which is: Naive-Bayes Algorithm, Support Vector Machine, Feed Forward Neural Network. There is no data on actual-time news and current model is run against the existing dataset, showing that the model performance depend on three factor i.e. the way model is trained, dataset used to train model and system configuration. We have projected a model for fake news detection via different machine learning techniques. In our future work, news article can be tested in real time scenario such as by creating browser extension so that we can minimize the clickbait and manipulation of larger population with false headlines.

References

[1] Datasets, Kaggle, February 2018.

[2] Y. Goldberg, "A Primer on Neural Network Models for Natural Language Processing," 2016 Journal of Artificial Intelligence Research (JAIR) Vol. 57(2016), pp. 345-420, doi: 10.1613/jair.4992.

[3] S. Afroz, M. Brennan and R. Greenstadt, "Detecting Hoaxes, Frauds, and Deception in Writing Style Online," 2012 IEEE Symposium on Security and Privacy, San Francisco, CA, 2012, pp. 461-475, doi: 10.1109/SP.2012.34.

[4] Hunt Allcott & Matthew Gentzkow, 2017. "Social Media and Fake News in the 2016 Election," Journal of Economic Perspectives, vol 31(2), pages 211-236, doi: 10.3386/w23089.

[5] Balmas M. When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism. Communication Research, 2014;41(3):430-454. doi: 10.1177/0093650212453600.

[6] A. Chakraborty, B. Paranjape, S. Kakarla and N. Ganguly, "Stop Clickbait: Detecting and preventing clickbaits in online news media," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, 2016, pp. 9-16, doi: 10.1109/ASONAM.2016.7752207.

[7] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14). JMLR.org, II–1188–II–1196.

[8] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436-44. doi: 10.1038/nature14539. PMID: 26017442.

[9] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[10] Z. H. Moe, T. San, M. M. Khin and H. M. Tin, "Comparison Of Naïve Bayes And Support Vector Machine Classifiers On Document Classification," 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), Nara, 2018, pp. 466-467, doi: 10.1109/GCCE.2018.8574785.