Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220029

Feed Forward Neural Network and K-Nearest Neighbour Based Comparative Accurate Social Media Spam Detection

Grandhi Sivadasu[a] and M. Adimoolam[b,1]
[a]*Research Scholar, Department of CSE, Saveetha School of Engineering,*
[b]*Associate Professor, Department of CSE, Saveetha School of Engineering,*
[a,b] *Saveetha Institute of Medical and Technical Sciences,*
[a,b]*Saveetha University, Chennai, Tamilnadu. India*

Abstract. The aim of this work is to perform spam detection in social media using Feed Forward Neural Network (FFNN) algorithm and compare its accuracy with K-Nearest Neighbour (KNN) algorithm. The experiment was carried out and classification was performed using KNN algorithm (N=10) for spam detection in social media and the accuracy was compared with SVM algorithm (N=10). For this experiment, G power value was calculated as 80 % and alpha value was as 0.05 %. The value obtained in terms of accuracy was identified for KNN algorithm (95.2%) and for FFNN algorithm (98.2%) with significant value 0.276. It was conclude that the accuracy of detecting spam using the FFNN algorithm give the impression to be slightly better than the KNN algorithm.

Keywords. K-Nearest neighbour, Spam detection, Social media, Feed forward neural network, Novel spam detection, Machine learning

1. Introduction

For spam detection in social media, some machine learning algorithms was proposed since a large amount of information has been misused. The spam detection was that it was the technique to detect spammer's information on the user's post content (Reddy and Srinivasa Reddy 2019) [1]. The application of detecting spam in social media is to detect the spam messages and identify the spam posts and who were misleading the customers. Thereafter performing feature extraction segment, nothing but tokenization development was applied to boundary the all-inclusive sentence into a group of words and hence extract the best features from the underdone data (Sangeetha, Nithyanantham and Jayanthi 2017) [2]. To choose a fitting worth of the separated list of capabilities, KNN has been pragmatic as an improvement scheming to adopt the ideal proficiencies from spam as well as non-spam information (Sultana et al. 2020) [3]. Spam detection techniques have a benefit that they use data to find high level features on their own, unlike the traditional machine learning algorithms (Crawford et al. 2015) [4]. Emails are utilized in most of the fields of education and business. These will be classified into ham and spam. The results of the experiments describes that the classification

[1]M. Adimoolam, Department of Computer Science and Engineering, Saveetha School of Engineering, Chennai, Tamilnadu. India. Email: adimoolamm.sse@saveetha.com

performance of KNN is high as compared with SVM rule and therefore the planned technique performance was saccharine as compared with the present state of the strategy (Talha and Kara 2017) [5]. While this paper concentrates on worries with email's spam, other spam structures truly do exist including interpersonal organization spam, blog spam, discussion spam and web index spam. This paper describes that how the spam was detected and predicted the trained values mentioned in the dataset with the machine learning algorithm (Ghiam 2012) [6].

One more work proposed for calculation using Multi-Layered Feed Forward Neural Network, which has a back-proliferation calculation (Bhowmick and Hazarika 2018) [7]. The information of the messages which were with the diagonals was extracted for the recognition purpose. This work got 92% accuracy of FFNN and 88% of KNN (Goyal, Chauhan and Parveen 2016) [8]. Last recreation results utilizing 10-overlap cross approval shows the best classifier in this work lessens the general blunder pace of the best model in the first paper referring to this dataset by the greater part (Firte, Lemnaru and Potolea 2010) [9]. It is important to add more messages to the dataset and trained dataset to detect the spam accurately (Ameen and Kaya 2018) [10]. As an author, the machine learning based price prediction solution has been carried out (John. et al. 2019) [11] and also carried out work related to COVID forecasting using machine learning algorithms and intelligent systems (Mohan et al. 2021) [12-15]. This research work was to predict spam detection in social media with better accuracy using FFNN and KNN algorithms and then compare these algorithms' accuracy.

2. Materials and methods

The research study was done in Saveetha School of Engineering. There are two groups considered in this research work: one is the KNN algorithm and the other group is FFNN algorithm. The sample size is iterated for 10 times with each algorithm and the pre-test power (G-power) is set as 80% and alpha value is set as 0.05 %. The dataset collected from the spam CSV. The Spam CSV dataset contains 50572 message records and datasets are a combination of ham and spam messages. The dataset is divided into two sets: Test and Train. Train set contains 80% ham and spam messages and the test set encloses 20% ham and spam messages (Ameen and Kaya 2018) [10]. KNN was taken as a gathering 1 calculation to identify spam and it is an apathetic learning, non-parametric calculation. It utilizes information with a few classes to anticipate the order of the new example point. KNN is not having parametric measures since it doesn't make any supposition on the data being contemplated and KNN model is upset from the data.

The working step of KNN is as follows.
Step-1: Start
Step-2: choose K numbers for neighbours
Step-3: Calculate Euclidean distance ED for K
Step-4: identify nearest neighbours for K based on ED.
Step-5: Count data point DP for K and iterate for each category.
Step-6: Assign new DP to calculate maximum for K.
Step-7: End.

FFNN was taken as a gathering 2 for identifying the most un-troublesome generous of neural association is a perceptron network with single layer, which contains a singular layer of yield center points; the information sources are dealt with clearly to the yields through a movement of burdens. How much the consequences of the heaps and the information not set in stone in each center point, and assuming that the value is over some edge (conventionally 0) the neuron blazes and takes the impelled worth (generally 1), else it revenues the disengaged regard (consistently - 1). Neurons with this thoughtful of motivation work are similarly called counterfeit neurons or direct breaking point units. In the composing the tenure perceptron often suggests networks comprising just one of these units.

Step-1: start
Step-2: set input numbers, say x
Step-2: take x to associate with social media data in single layer, for example=f*(x)
Step-3: Map the inputs like Y=f(x; 0) and learn the value of the parameter theta
Step-4: The data flows through the function being evaluated from x
Step-5: The index based on input will execute the output.
Step-6: The intermediate computations used to define f and finally to the output
Step-7: Model is ready

This study is worked on a system with Intel i5 core processor using 8 GB ram and uses a ROM of 1TB. This system is Windows OS as its base operating system. It is recommended to install anaconda navigator to launch Jupyter notebook. This review involves the test T as its component. A sort of inferential as test and static cast-off was adopting further. There is a massive is similarity amongst the mean of two gatherings. It was strength to ally with precise elements. This experiment is utilized to improve exactness of the specific calculation. These experiments are likewise utilized in theory testing insights. From this analysis, it is fundamental to get an instrument with Novel Spam Detection. The statistical tool used for this study is IBM SPSS version 21. The independent variable is message length and dependent variables are positive message and negative messages. In SPSS the dataset is prepared using 10 iterations as sample size for the both classifier FFNN and KNN algorithms.

3. Results

Both FFNN and KNN algorithms' accuracy values are measured 10 times and these predicted values are listed with standard deviation and standard error mean as Table 1.

Table 1. T-test prediction information of KNN and FFNN algorithms' dataset group statistics for mean accuracy, std. Deviation and std. Error means details

Group	Algorithm	Mean	N	Std. Deviation	Std. error mean
Accuracy	KNN	84.0910	10	8.29514	2.3069
	FFNN	87.0984	10	8.4401	1.5063
Precision	KNN	92	10	7.8925	2.3401
	FFNN	95	10	8.5401	1.8520

The mean, standard deviation and standard mean error rate has been inserted. The FFNN algorithm has a lower error rate than the KNN algorithm. KNN algorithm has

mean accuracy of 84.098 % and FFNN algorithm has mean accuracy of 87.091 % for the epoch value of 10. The significance value is calculated and shown in Table 2.

Table 2. Mean accuracy, mean difference, levene's test for likeness of positions, 95% confidence interval of difference, standard difference of error with significance value 0.276

	Levene's test for equality of variables		**T-test for equality of means**					**95%confidence interval of the difference**	
	F	sig	t	df	Sig (2 tailed)	Mean difference	Std. error difference	Lower	Upper
Accuracy	1.2	0.276	1.77	18	0.093	6.71	3.78	-1.22	14.6
Accuracy			1.77	16.89	0.094	6.71	3.78	-1.26	14.6

It is experimental and difference of significant amongst the two algorithms was also measured and significance value gained as 0.276. FFNN algorithm shows the effective results than the KNN algorithm considerably, and the mean difference and standard error difference is tabulated in Table 3.

Table 3. How the spam messages are detected using KNN and recreating the original form of the message

label	**Message**	**len**	**clean_message**	**Clean message len**	**label**
0	Available only... Go pending point...	111	Go jurongfactcrazigain n prodigiousecosphere	76	1
1	Ok lar...was flippant with u oni...	29	Ok laranecdotewif u oni	21	1
2	Free entry in 2 a wkly comp to win FA Cup final	155	Ablewkli comp win fa cup finishingcouponst m...	99	0
3	U dun roughlyprimary hor...u c previouslybeforearound....	49	U dun roughlyeatlihor u c previouslyroughly	35	1
4	Nah i don't reflect he enthusiasms to usf,he subsistsaro..	61	Nah deliberateenthusiasmsusf around through	36	1

The spam messages are detected receiving from the spammer using the FFNN algorithm. It was detected the spam using stop words method and report it as spam message. Initially the dataset contains several messages which include spam and ham messages. 80% messages are taken for training and 20% of messages are taken for testing. The length of the spam and ham messages present in the dataset and it was detected the spam according to the given conditions and length of the messages accordingly. The FFNN algorithm is used to detect the spam messages getting from the spammers with the help of this algorithm. This proposed algorithm can easily detect spam with better accuracy as shown in Figure 1. The detailed analysis of the dataset and data balancing of taken dataset was carried out. Detecting the spam letters from the dataset is shown using the word cloud.

Comparison of FFNN and KNN algorithms including error rate for the sample size of 5572 and the epoch value of 10 was carried out and these results are shown in Figure 2. The standard deviations among the algorithms are slightly different. X Axis: FFNN versus KNN algorithms. Y Axis is a mean accuracy comparison of FFNN and KNN algorithms ± 1 SD. The compared results conclude that FFNN has higher accuracy and

lesser error rate than the KNN algorithm. Thus FFNN implementation of this research is introducing a Novel Spam Detection.

Table 4. How the spam detected from the inbox send by the spammer using FFNN.

	label	Message	len	Clean message	Clean message len
0	ham	Go until point...available only...	111	Go jurongargumentcrazi purpose n unlimitedcreation	76
1	ham	Ok lar...joking with u oni...	29	Ok largagwif u oni	21
2	ham	Permittedentrance in 2 a wkly comp to landslide FA Cup ultimate	155	Free wkli comp win fateacupfinishingper mitst m...	99
3	ham	U dun say initial hor...u c previouslyformerlyro ughly....	49	U dun aroundeatlihor u c previouslyroughly	35
4	ham	Nah i don't contemplate he drives to usf,he survivesaro..	61	Nah reasonenthusiasmsusf everywherecomplete	36

Figure 1. How word cloud helps to detect the spam messages from the experimented dataset.

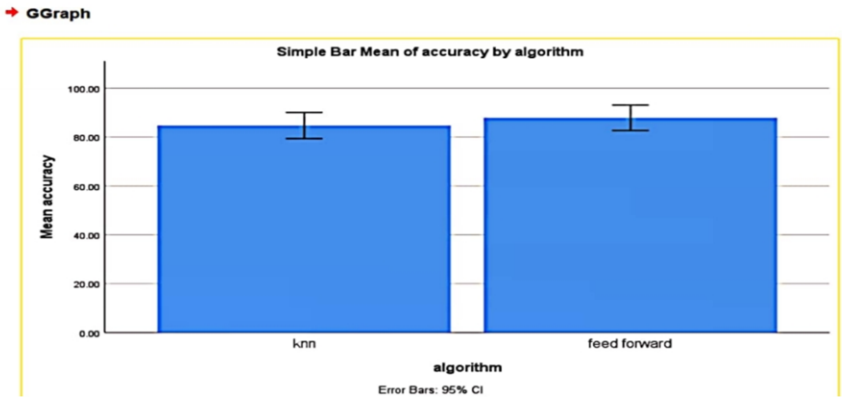

Figure 2. Bar graph for comparing the mean accuracy (87%) using FFNN and KNN algorithms.

4. Discussion

The proposed work analyzed the accuracy and precision of FFNN and KNN algorithms. The result shows the evidence that there is statistical difference between the FFNN algorithms (98.2%) and KNN algorithm (95.2%). The FFNN algorithm accuracy appeared to be higher than the KNN algorithm and compared accuracy with existing work (Suleiman and Al-Naymat 2017) [16]. The work describes how work has detected the spam and the difficulties rose, while detecting the spam in social media. The SVM algorithm is mostly used to classify text-based brochures (R. Kumar, Ghadage and Navale 2014) [17]. Though KNN algorithm is a record cast-off practice in article cataloguing its routine in the junk recognition is not best when compared to FFNN algorithm. This system would surge the concert by unravelling the delinquent of circulation of working out data (Sasaki and Shinnou 2005) [18].

One research work described spam detection systems and showed their results using a particular algorithm with less accuracy. The method was tested on the dataset of the benchmark. One drawback of these proposed classification algorithms is that accuracy is not high since limited iteration has been done with limited features. Hence classification might not be accurate in the detection of spam and time consumption of the model is high due to complexity in the algorithm. The major problem is the email's messages and it is an unwanted message (spam). The individual that sends the spam messages is known as a spammer who gathers email addresses from sites, talk rooms, and infections. The algorithms in this current work have not been able to remove the requirements of manual checking of the spam detection. So this limitation will be removed in future work for systems with maximum efficiency as the usage of social media. The online stores will also increase so it has to increase the accuracy to detect the spam messages sent by the spammers. This is necessary in order to produce a counter algorithm.

5. Conclusion

It has been concluded from our study that the FFNN algorithm accuracy (98.2%) and the FFNN algorithm appeared to perform more effectively and accurately than the KNN algorithm accuracy (95.2%). The accuracy of the FFNN algorithm is a good score in predicting the spam detection in social media.

Reference

[1] Reddy Kurapati Subba, E. Srinivasa Reddy. Integrated Approach to Detect Spam in Social Media Networks Using Hybrid Features. International Journal of Electrical and Computer Engineering. 2019 Feb;9(1):562-569.

[2] Sangeetha M, S Nithyanantham, MJayanthi. Comparison of Twitter Spam Detection Using Various Machine Learning Algorithms. International Journal of Engineering & Technology. 2018;7(1.3):61-65.

[3] Sultana Thashina, KA Sapnaz, Fathima Sana, Jamedar Najath. Email Based Spam Detection. International Journal of Engineering Research & Technology. 2020 Jun;9(6):13-139.

[4] Crawford Michael, Taghi M Khoshgoftaar, Joseph DPrusa, Aaron N Richter, Hamzah Al Najada. Survey of Review Spam Detection Using Machine Learning Techniques. Journal of Big Data. 2015 Oct;2(1):1–24.

[5] Talha Abdullah, Resul Kara. A Survey of Spam Detection Methods on Twitter. International Journal of Advanced Computer Science and Applications. 2017 Mar;8(3):30-38.
[6] Ghiam Shekoofeh. A Survey on Web Spam Detection Methods: Taxonomy. International Journal of Network Security & Its Applications. 2012 Sep; 4(5):119-134.
[7] Bhowmick Alexy, Shyamanta M Hazarika. E-Mail Spam Filtering: A Review of Techniques and Trends. Lecture Notes in Electrical Engineering. 2017Oct;443:583-590.
[8] Goyal Saumya, R K Chauhan, Shabnam Parveen. Spam Detection Using KNN and Decision Tree Mechanism in Social Network. Proceedings of Fourth International Conference on Parallel, Distributed and Grid Computing; 2016 Dec 22-24; Waknaught, India: p. 522-526.
[9] Firte Loredana, Camelia Lemnaru, Rodica Potolea. Spam Detection Filter Using KNN Algorithm and Resampling. Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing; 2010 Oct 21; Cluj-Napoca, Romania: p. 27-33.
[10] Ameen Aso Khaleel, Buket Kaya. Spam Detection in Online Social Networks by Deep Learning. International Conference on Artificial Intelligence and Data Processing. 2018 Sep. 28-30; Malatya, Turkey: p. 1-4.
[11] John A, D Praveen Dominic, M. Adimoolam, N M Balamurugan. Share Market Data Prediction Strategies Using Deep Learning Algorithm. Recent Advances in Computer Science and Communications. 2019 Dec;14(6):1852-1859.
[12] M Adimoolam, Karthi G, A John, Mohan Senthilkumar, Ahmadian Ali, Ciano Tiziana. A Hybrid Learning Approach for the Stage-wise Classification and Prediction of COVID-19 X-ray Images. Expert Systems.2021 Dec;39(4):1-15.
[13] Balamurugan N M, T K Rathish babu, M Adimoolam, A John. A Novel Efficient Algorithm for Duplicate Video Comparison in Surveillance Video Storage Systems. Journal of Ambient Intelligence and Humanized Computing. 2021 Apr;1(1):1-15.
[14] Mohan Senthilkumar, A John, Ahed Abugabah, M Adimoolam, Shubham Kumar Singh, Ali Kashif Bashir, Louis Sanzogni. An Approach to Forecast Impact of Covid-19 Using Supervised Machine Learning Model. Software: Practice and Experience. 2021 Apr;52(4):824-840.
[15] John A, T Ananth Kumar, M Adimoolam, Angelin Blessy. Energy Management and Monitoring Using IoT with CupCarbon Platform. Green Computing in Smart Cities: Simulation and Techniques. 2021:1-24.
[16] Suleiman Dima, Ghazi Al-Naymat. SMS Spam Detection Using H2O Framework. Procedia Computer Science. 2017;113:154-161.
[17] Kumar Ritesh, Shital Ghadage, GS Navale. Spam Detection Using Approach of Data Mining for Social Networking Sites. International Journal of Computer Applications. 2014;108(9):16-18.
[18] Sasaki, M., and H. Shinnou. 2005. “Spam Detection Using Text Clustering.” 2005 International Conference on Cyberworlds (CW’05). https://doi.org/10.1109/cw.2005.83.