Fake site detection using Machine Learning Algorithm and N-Gram Analysis

Asha J[1] , Saradha S[2], Mohanraj A[3], Devipriya C[4]

ashaantony2805@gmail.com[1], saradha.s@sece.ac.in[2], mohanraj.a@sece.ac.in[3]

PSG Institute of Technology and Applied Research[1], Sri Eshwar College of Engineering[2,3]

Abstract. The prevalence of fake websites increases as more people uses the internet. The identification of fake internet sites has thus been the subject of increased research in recent years.Website detection is extremely difficult because there aren't enough resources, or datasets available. This study seeks to identify fake websites by integrating two independent feature extraction techniques with Support vector machines, KNN, and logistic regression as learning algorithms. Term frequency (TF) and Term frequency-invented document frequency (TF-IDF) using N-gram analysis is also deteremined. To support the research, performance metrics are analyzed including accuracy, precision, and recall. The experiment results in a high TF feature extraction accuracy when N = 1, and a high TF-IDF feature extraction accuracy when employing the KNN algorithm (97.3%).

Keywords: Fake sites, feature extraction, N-gram analysis, SVM, KNN, LR

1 Introduction

Internet use is very important in underdeveloped countries. As the number of internet users grows, so does the dissemination of false information. (That is, the distribution of false information is closely correlated with internet usage). Real-world examples include the majority of unemployed youth who applied via a fraudulent website and wasted their time and money. Although there are more real-world instances that may be utilized to show bogus websites, the proposed technique aids in avoiding this kind of fraudulent websites, which include various misleading giveaways and security alerts with the primary purpose of pretending to be authentic websites. [1] The numerous methods in which a phoney website manipulates consumers through emotional manipulation by sending dubious signals. The first one is the type of urgency messages which push you to do the immediate action to do before thinking the situation. The second sort of excitement is the most concentrated on attracting the customer by delivering vouchers and free gift cards, which prompts the next motions. The third category is something critical that instills dread in the user through false virus and security notifications. Internet scams are unlikely to slow down as we transition to a new normal; instead, they are likely to worsen.

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343171

Knowing how to verify the credibility of a website will help you avoid problems caused by bogus websites both now and in the future. While 80% of problems can be solved by effective problem identification, the remaining 20% may still exist. Easy methods for inspecting websites include checking for misspelled URLs, looking for site seals and locks, comparing secure websites with scam ones, looking beyond the locks, and running the website via a website checker. Check to see whether the URL is misspelled; a misspelled URL, as well as more space and special characters, is a crucial signal of a bogus site. Fraudsters may subtly alter a URL name, such as using xyz.com, or they may modify the domain extension, such as xyz.org instead of xyz.com. Businesses such as Facebook, Twitter, TikTok, Google, Pinterest, Tencent, and YouTube are collaborating with WHO to minimize the spread of rumors. Their efforts are aimed at deleting information that may be hazardous to the public's health. Datasets are used to improve algorithms. These datasets can be separated into two categories: training and testing. In many of the studies I've seen, a system combines several machine learning algorithms with data mining. This is common on social networking sites, especially with Twitter data. There are numerous methods to assist in this conflict. However, it is crucial to comprehend the various methods for false news identification that are already in use before moving forward. We'll look at it from both manual and automatic perspectives. Search for site seals. A seal of the websiteexpresses that thesite is real and unprotected, and you may usually click on it to see further information about the website, such as information that should be kept secret and how the website's validity was certified. Clicking on a seal that does nothing should not be believed because it is most likely a forgery.A padlock on a website does not mean that it is secure or that it is not a fraud. According to studies, a padlock is already used on up to 50% of phishing websites. [2] A padlock on a website signifies that the data is encrypted and thus safe from browsers' perspectives. In today's world, however, a protected website does not indicate that it is safe to use for a secure transactions or information exchange. Although phony websites that employ several layers of transport ans socket certificates are routinely blocked, they may be able to temporarily disrupt a certificate. Internet data is encrypted using Transport Layer Security (TLS) to prevent hackers and eavesdroppers from accessing what the users send. This is especially helpful for sensitive and private information like credit card particulars, passwords, and correspondence.

2 Literature Survey

Nawafleh and Hadi [3] used a novelclassification technique to identify spam sites. According to the results of an observational study, successful people are more likely to know one other. When compared against various calculations, categorization has been proved to be competitive, as demonstrated by instances such as PRISM, RIPPER, and NB.

The tactic [4] Using a random forest algorithm, PILFER (Phishing Identification by Learning on Features of Email Received) properly recognised 96% of phishing communications with a 0.1 percent false-positive rate. A variety of learning techniques, including SVM, decision trees, rule-based methods, and PILFER, were tested by the researchers. The trial findings take into account the following email components: HTML messages, Number of Links, Domains, Dots, and Number of Links. IP address URLs, Domain Age, Non-Coordinating URLs, "Here" Link, and Non-Coordinating URLs.

Spam-channel The Recommended approach [5] suggested a model for figuring out whether a website is phishing or not, and the output includes a Java script. With relative classification accuracy scores of 92.7846 percent, 95.11 percent, 96.57 percent, 96.3 percent, and 93.85 percent, they use six different machine learning-based categorization algorithms. Among these methods are Nave Bayes, J48, SVM, Random Forrest, Tree Bag, and IBK lazy classifier. For them, testing receives 20% of the budget and training receives 70%, a ratio of 70 to 30. In this experiment, we extend their work and utilise those techniques to improve classification accuracy. We introduce a revolutionary classification system called Neural Net.

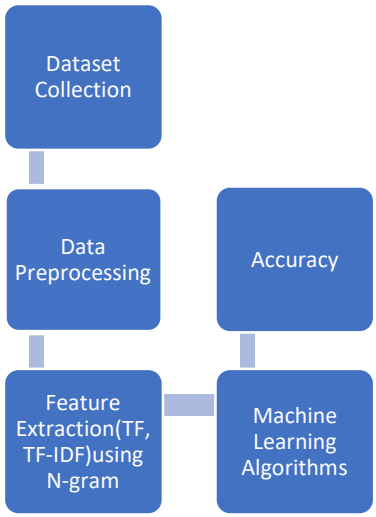
The technique demonstrates how to identify bogus news utilising feature extraction methods including term frequency (TF) and word frequency, as well as machine learning algorithms (TF). Six machine learning methods are compared with inverse document frequency, with SVM, logistic regression, and stochastic gradient descent providing the most accurate results. False news was discovered in a manner akin to how fraudulent websites were discovered [6].

By using feature selection in the context of phishing location, Bergholz et al. [7] demonstrated how to enhance learning models for detecting phishing communications. Using the classifier as a feature of the evaluation function and picking a subset of components using a wrapper technique, the alleged best-first pursuit calculation swiftly adds and subtracts features from an existing subset.

3 Proposed Method

The proposed research makes use of machine learning techniques, and helps in the identification of fraudulent

Fig. 1. Proposed Method



websites. This research utilizes a dataset that is readily available. Once the data is collected, pre-processing of the same is carried out as the first and foremost step. This would help in improving the accuracy of the dataset. Fig 1 shows the steps involved in our proposed approach.

The Acquired data is pre-processed in order to prepare it for additional analysis or primary processing. Pre-processing techniques could include retrieving data from a larger set, filtering it for a variety of reasons, and combining data sets. Data purification, data transformation, and feature selection are the steps involved in the pre-processing of data.

3.1 Feature extraction

When you need to analyse fewer resources without losing any significant or meaningful data, feature extraction can be helpful. Also, it helps to reduce the amount of duplicated data required for feature analysis. Words are used to build features, which enriches the corpus's context and broadens the scope of the characteristics. Machine learning algorithms study a predetermined set of features from the training data to provide results for the test data.However, the main problem with language processing is that machine learning methods cannot directly handle text. We will require a variety of feature extraction approaches to convert text into a matrix or vector of features.

3.2 N-Gram model

A continuous run of n items from a particular text or audio sample is known as an N-gram. [8] The dataset could be tokens, phonemes, syllables, letters, words, or base pairs, depending on the application. From a corpus of text or voice, n-grams are frequently extracted. Natural language processing techniques and text mining both make extensive use of text's N-gram. In essence, they are a collection of words that appear together in a particular window and are calculated to promote a single word. The word count is represented by the number N, where n=1 denotes a single word, n=2 a double word, and so on.

3.3 Term Frequency

The phrase's frequency in a document (TF) is the number of times it appears. In natural language, words and sentences can be substituted for one another. However, a word can stand in for any text token. Paper length varies, therefore it's possible that a phrase will come up more often in longer papers than in shorter ones. A sentence will therefore have greater weight in a longer document than in a shorter one. To normalize the effect, term frequency is frequently divided by the total number of terms in the document.

3.4 Term Frequency -Inverse Document Frequency

A measure of how significant a word is to a document in a corpus or collection of documents is inverse document frequency, a numerical statistic that weighs an uncommon period in a group of documents. [9] Finally, the weight of rare terms is determined by term frequency and inverse document frequency (TF-IDF), which count all phrases together. Uncommon terms or keywords in publications have been discovered to have a higher level of relevance.

3.5 Machine Learning Algorithms

Machine learning is the process through which computers learn to accomplish things without being explicitly taught. Computers use data to learn how to perform specific tasks[10]. For simple tasks handed to computers, there is no need for the computer to learn; instead, it is simple to create algorithms that explain to the machine how to carry out all steps required to address the problem at hand. Python's ease of use allows developers to construct dependable solutions,

whereas machine learning and AI rely on sophisticated algorithms and adaptive procedures. Python code is simple for humans to understand, which makes it easier to create machine learning models. The classification procedure employs a machine learning approach to train the classifier and evaluate accuracy.

3.5.1 Support Vector Machine

A linear model called the Support Vector Machine (SVM) can be used to solve regression and classification issues. It can handle both linear and nonlinear problems and is useful for many different applications. [11] SVM is a simple concept that anyone can comprehend.The program first creates a line or a hyperplane to categorize the data.Step 2: Two categories are separated by a hyperplane, with the distance of the nearest feature to the hyperplane being taken into consideration.The technique identifies the points in both classes that are most near the line. These points are referred to as support vectors. The line's separation from the support vectors is now calculated. [12] The margin is the name given to this distance. The optimal hyperplane is the one with the greatest margin. An SVM model is used for relating points in space, with a minimum distance as feasible between the examples of each category.SVMs effectively perform both linear and non-linear classification by implicitly transforming their inputs into sizable feature spaces..

3.5.2 K-Nearest Neighbors

The KNN methodology is the widely utilised learning algorithm for the classification strategies. KNN is a slow, non-parametric learning method. Theaim of K Nearest Neighbours algorithms is to predict the classification of a new samples point using a collection of information elements that are categorised into various classes [13]. K nearest neighbours is a straightforward strategy for classifying new cases based on a similarity score while retaining all current examples (such as distance functions).[14] Pattern recognition and statistical estimation have both made use of KNN.

Step 1: Import the training dataset and calculate the distance between each point in it and the item to be categorized.

Step 2: Select the locations that are closest to one another and have the K smallest distance between them.

The third step is to hold a "majority vote."

3.5.3 Logistic Regression

Logistic regression is a straightforward strategy that is straightforward to grasp and execute. It can reveal the cooridnation between the known variables and the likelihood of a yeild the appropriate result. The relationship between the categorical unknown variable and one or more independent variables is computed using logistic regression. [15]. The cumulative distribution is a special technique used in logistic regression to calculate probabilities using a logistic function. Learning against likelihood is frustrating, but parameter estimation gives better results that are unbiased and have less variance. Independently calculate hypothesis ratios, log functions and relative probability.The model is known as a multiple or multivariable logistic regression model when there are several variables (for example, risk factors and treatments), and it is one of the most often used statistical models in medical publications. This chapter looks

at categorical, continuous, and multiple binary logistic regression models, as well as interaction, quality of fit, categorical predictor variables, and multiple predictor variables.

4 Experiment Result

4.1 Software Requirements

The cloud environment is a Jypter notebook with GPUs and TPUs for intense processing and a Windows operating system. Python was the programming language utilised. As a result, a few public datasets are available. This project makes use of a new dataset from Kaggle. Set includes more trustworthy data and less inaccurate site reports. The experiment's dataset is divided into 80% training te model and 20% were used to testing groups.

4.2 Implementation

The three aforementioned techniques were utilised to create learning models (from the training dataset), which were afterwards applied to predict the labels given to the testing data. The experiment's findings were then reviewed, analysed, and interpreted. The two alternative feature extraction approaches, TF-IDF and TF, were used, with n-gram sizes ranging from 1 to 3. Table 1 shows the performance of support vector machine. N-Gram ranges from 1, 2 and 3 for TF and TF-IDF is evaluated and the performance is shown in table.

Table 1.Performance of SVM (in terms of Accuracy)

N-Gram	TF	TF-IDF
1	88.4	83.1
2	74.2	74.2
3	67.4	67.1

Table 2. PERFORMANCE of KNN (Accuracy in terms of %)

N-Gram	TF	TF-IDF
1	98.7	97.3
2	81.5	78.5
3	76.4	72.4

Table 2 depicts the performance of KNN by varying the N-Gram values from 1, 2 and 3 is evaluated and the performance is shown in the table. Table 3 represents the performance of LR for various N-Gram in terms of TF and TF-IDF.

Table 3. PERFORMANCE of LR (in terms of Accuracy)

N-Gram	TF	TF-IDF
1	97.1	91.4
2	94.5	90.5
3	91.6	83.4

5 Performance Evaluation

Performance evaluation based on the accuracy of several Machine Learning algorithms is shown in the Tables[1, 2 ,3]. Fig. 2 and 3 shows the analysis of performance in terms of graphical representation. Performance analysis in terms of Term Frequency is shown in Figure 2. Performance of N-Gram (N=1 for TF) using SVM obtained 88.4%, KNN obtained 98.7% and LR achieved 97.1.

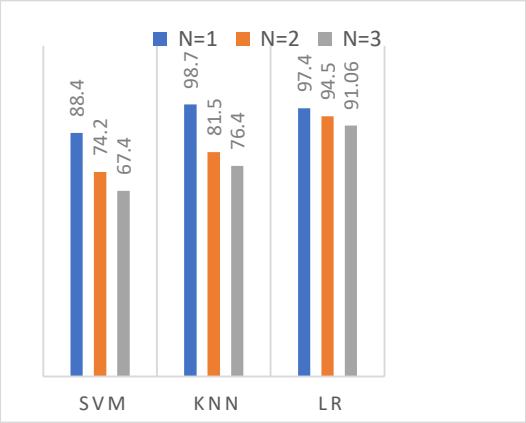


Fig. 2 : TF analysis for various ML Technique

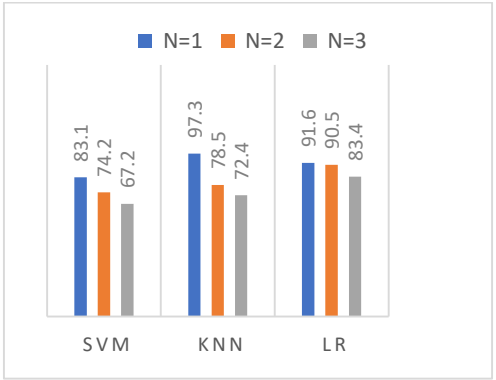


Fig. 3 : TF-IDF analysis for various ML Technique

Performance analysis in terms of IDF is shown in Figure 3. Performance of N-Gram (N=1 for IDF) using SVM obtained 83.1%, KNN obtained 97.3% and LR achieved 91.4.

When the N value gets increases accuracy level of each algorithm get decrease. Therefore, applying n-gram analysis on fake site detection is not so effective. From, Figure 3 the results obtained in our experiment for TF-ID was same stage for both KNN and LR. In TF-IDF analysis accuracy level is less when compared to TF analysis, Therefore, choosing TF as a feature extraction technique is best for fake site detection specifically for predefined dataset. From the analysis it is observed that KNN using 1-Gram provides highest accuracy of 98.7% than SVM and LR.From[13] As is evident from the experiment results, classifier accuracy rises as the amount of training data increases. The proposed methodology with the ratio 34 out of 66 for Random Under-sampling is summarized in Table 6 list in the paper [13]. In this fraction, both Naive Bayes and logistic regression have a specificity rate of 1.0, meaning that they both correctly classified the non-fraud cases. This may be the case because as the training data sample grows, so does the accuracy of both classifiers. Since both techniques estimate prioriprobability, which rises with the number of samples in the training data, they aid in more accurate classification of the data samples.

6 Conclusion

In this paper, the investigation of unique machine learning-based strategy for classifying bogus news. The creation of a technique that allows the TF-IDF Vectorizer to differentiate between authentic and fake news is discussed in this study. Kaggle datasets are utilized during implementation. The outcomes show that this technique works well.N-gram analysis was utilized to create a detection model for bogus sites using several feature extraction approaches and machine learning algorithms. When employing unigram features extraction as term frequency (N=1) and KNN classifier (Machine learning Algorithms), the proposed methodology achieves the maximum accuracy. The project can be discussed further with some deep learning methods.

References

[1] Qbeitah, M.A. and Aldwairi, M.,: Dynamic malware analysis of phishing emails,9 th International Conference on Information and Communication Systems (ICICS), pp. 18-24 (2018).

[2] Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E.,: Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In Proceedings of the 3rd symposium on Usable privacy and security, pp.88-99 (2007).

[3] Nawafleh, S. and Hadi, W.E.,: Multi-class associative classification to predicting phishing websites. International Journal of Academic Research Part A, 4(6), pp.68-73 (2012).

[4] Fette, I., Sadeh, N. and Tomasic, A.,: Learning to detect phishing emails.,Proceedings of the 16 th international conference on World Wide Web, pp. 649-656 (2007).

[5] Tiwari, P. and Singh, R.R., : Machine learning based phishing website detection system, International Journal of Engineering and Research, 4, 12, 2007.

[6] Asha, J. and Meenakowshalya, A.,: Fake News Detection Using N-Gram Analysis and Machine Learning Algorithms. Journal of Mobile Computing, Communications & Mobile Networks, 8(1), pp.33-43 (2021).

[7] Islam, M. and Chowdhury, N.K., November. Phishing websites detection using machine learning based classification techniques, International Conference on Advanced Information and Communication Technology, Chittagong, Bangladesh, 2016.
[8] Ahmed, H., Traore, I. and Saad, S.,: October. Detection of online fake news using n-gram analysis and machine learning techniques. In International conference on intelligent, secure, and dependable systems in distributed and cloud environments, pp. 127-138). Springer, Cham, (2017).
[9] Liu, Q., Wang, J., Zhang, D., Yang, Y. and Wang, N., 2018, December. Text features extraction based on TF-IDF associating semantic, IEEE 4th International Conference on Computer and Communications (ICCC), IEEE , pp. 2338-2343(2018).
[10] Haldorai, A. and Kandaswamy, U.,: Supervised Machine Learning Techniques in Intelligent Network Handovers. In Intelligent Spectrum Handovers in Cognitive Radio Networks, pp. 135-154,Springer, Cham (2020).
[11] Praveena, M. and Jaiganesh, V.,: A literature review on supervised machine learning algorithms and boosting process. International Journal of Computer Applications, 169(8), pp.32-35(2017).
[12] Devikanniga, D., Ramu, A. and Haldorai, A.,: Efficient diagnosis of liver disease using support vector machine optimized with crows search algorithm. EAI Endorsed Transactions on Energy Web, 7(29), (2020).
[13] Itoo, F. and Singh, S.,: Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13(4), pp.1503-1511, (2021).
[14] Saranya, N., Samyuktha, M.S., Isaac, S. and Subhanki, B., March. Diagnosing chronic kidney disease using KNN algorithm, 7 th International Conference on Advanced Computing and Communication Systems (ICACCS), Vol. 1, pp. 2038-2041, 2021.
[15] Radhika, P.R., Nair, R.A. and Veena, G., 2019, February. A comparative study of lung cancer detection using machine learning algorithms, IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1-4, (2019).