Lexicon and Learn-Based Sentiment Analysis for Web Spam Detection

Anum Waheed, Dr Abdus Salam, Dr Javed Iqbal Bangash, Maria Bangash

[1](1st Affiliation) Department of Computing and Technology, Abasyn University, Peshawar, Pakistan; [2](2nd Affiliation) Department of Computing and Technology, Abasyn University, Peshawar, Pakistan; Institue of Computer Science and IT, Peshawar, Pakistan; Department of Computing and Technology, Abasyn University, Peshawar, Pakistan.
Email: address desired (anumwaheed@outlook.com)

ABSTRACT

The increased usage of internet has also made it an opportunity for cybercrimes and unethical activities. Spam and fake news propaganda are critical aspects in this domain. Spam and fake news have adverse impacts on everyone and can cause serious negative outcomes. Spam detection has therefore gained much attention in this era. Catering to this dilemma, this study aims to develop models for spam and fake news detection so that the spam and fake news may be identified and removed from internet. For this purpose, a hybrid sentiment analysis approach is utilized and state of the art machine learning algorithms like Naïve Bayes, Random Forest and RCNN are developed for fake news detection. The data is collected from different news websites i.e. 700,000 news articles, and another dataset includes Kaggle dataset of news articles i.e. 6256 in number. A web scrapper tool is developed with multiple criteria for deeming websites fake or authentic to extract useful information from dataset. Preprocessing of the data is carried out i.e. stemming, tokenization and removing stop words. The term frequency and document term matrix are developed and the developed models are trained and tested with datasets. The three mentioned models are used for data analysis for both datasets. The results show that RCNN algorithm in hybrid sentiment analysis approach is the most suitable, effective and accurate model for spam and fake news detection. With the application of this model and method, the fake news can be detected and removed making the internet more reliable and secure.

Keywords : Sentiment analysis, Fake news detection, Spam, Machine learning algorithms, Random Forest, RCNN, Naïve Bayes

1 INTRODUCTION

1.1. Background

The advances in technology have enchanted implications in all spheres of life. The internet boom and developments in internet related technologies are embraced by everyone in this modern era and have also become crucial part of everyone's daily lives [1]. Many necessary experiences and activities have been entailed in the usage of internet which includes entertainment, education, shopping and communication and so on [2]. MacDermott et al. [3] explained that the dilemma is that when technologies are adopted by the people, the negative impacts also arise alongside the positive ones. As the online activities and online life is improving and increasing for everyone so are the cybercrimes and criminal or unethical activities. The internet has also become an opportunity for the unethical and criminal activities because much data is available on the internet and it can be manipulated by the negative minds for personal favours and others' losses. The crimes related to internet include phishing, spamming, spreading fake and wrong news and information, harassing, identity theft, misusing personal information and hacking [4].

Among them, the spread of false and fake news is critically important to be tackled and explored. This is because fake news and wrong information is generated and spread across internet reaching out to millions of people who can believe them and make decisions accordingly [5]. This process called spamming, where people use false information and fake news and send it to multiple people for personal benefits, illicit purposes, illegal activities, commercial advertising, spreading malware and phishing, can generate serious consequences for the end users. For example, such spams can cause collective hysteria, terrorism, deceptive thoughts, forced opinions and wrongful usage of information by the perpetrators [6].

Spam detection and mitigation is a complex phenomenon because of prevalence of abundance of fake websites, fraudulent and false news, SMS, social communication and emails platforms [7]. Spam websites also aim to deceive search engines to enhance their rankings and become the reliable source of information on search engines apparently to the end users [8]. Therefore, tackling such complicated and critical

problem requires holistic approach. In this paradigm, sentiment analysis or opinion mining is a promising phenomenon that can be used for spam detection and mitigation because this data processing involves text data analysis [9]. Most of the data on the internet is in the form of text and therefore, spam detection of this data using sentiment analysis is a practical approach [10]. Asghar et al. [11] explained that sentiment is explained as the personal or exaggerated judgment or belief of something that is not based on proof or certainty. Mahajan and Rana [12] argued that in this technical and social era, the design and development of technical tools which can tackle abusive words, spam detection and proper sentiments of text for public opinion studies and data mining can prove very practical and beneficial. Therefore, this research papers aims to utilize sentiment analysis approach for the detection of spam websites and fake news and validate the methodology adopted so that the results might be deemed practical and reliable.

1.2. Problem Statement

The problem statement for this research paper entails the spam news that can generate misunderstanding and inconvenience for the users on internet and the complexity of tackling these spam news i.e. spam detection and mitigation. Such fake news or spam can cause harmful impacts on people's lives. Therefore, tackling spam and fake news is crucial and the detection of spam websites and fake news data based on the sentiment analysis approach and also the validation of the proposed methodology to analyze the accuracy of the results. Spam detection requires technical approaches to be handled and sentiment analysis method can prove useful but the validity of the algorithms used in sentiment analysis is challenging [10]. Therefore, this problem is solved in this research paper because the spam detection, utilization of sentiment analysis and validation of accuracy of the results are all tackled.

1.3. Objectives

Following objectives are addressed via this research:

- To classify the spam and ham websites of news based on sentiment analysis
- To develop the sentiment analysis methodology based on hybrid method and web scrapper tool for collecting and filtering news' websites contents
- To validate the results and proposed method by analyzing the accuracy of different sentiment analysis algorithms and comparing them for different datasets and methods

Kaggle is a data science company that provides access to public datasets and data analysis tools [13].

1.4. Scope

The scope of this research entails the fake news websites and fake news data. Moreover, the sentiment analysis method developed for this research includes hybrid approach. The data collection is based on different news websites and results along with proposed methodology are validated for accuracy by using Kaggle datasets too. Other spam websites and spam types are not included in the scope of this research.

1.5. Significance

The key research areas in internet and social network include spam detection and sentiment analysis. Much research work has been conducted in these areas but limited to classifying fake news and fake websites. In contrast, there is an absence of a collective and well-organized technique that can tackle the dynamics of these areas i.e. spam detection and fake news and websites detection. This research aims to put efforts in this direction where the development of an effective technique for spam detection using combined sentiment analysis approach and web scrapper and Kaggle datasets would lead to more accurate results. Internet is the basic necessity in almost all of the fields in life and is used by most of the people. Therefore, detecting spam and removing it from internet is crucial so that true information may become accessible to everyone. In addition, one of the best methods of machine learning classification algorithms i.e. Python libraries, is used in the methodology that identifies the fake news and websites with more accuracy.

1.6. Structure

The research paper first introduces the research and then literature review is presented in the second section. After literature review, methodology developed for this research is presented in the third section. The fourth section presents the results and discussions and in the end, conclusion is presented.

2 Literature Review

Agogo and Hess [14] argued that emergence of new technologies and opportunities in internet world have brought about many unethical and criminal activities that can harm people seriously. These activities include cybercrimes related to harassing, phishing, spamming, financial crimes, identity theft and so on. One notable issue in these crimes is spread of fake news and spam across internet among millions of people [15]. This is because of high prevalence of internet usage worldwide and therefore, such spams and fake news impact millions of people [5]. Fake websites, fake news and spams have become abundant in the internet domain and people are losing trust over internet for the news. For example, as per research by Watson [16], 52% of the people in the United States believe that the news they hear on internet are fake. This dilemma is further extended because there are many types of internet related crimes and unethical activities and many types of fake websites are involved in such activities.

Carpineto and Romano [17] classified fake websites into three major types which include web spams, concocted websites and spoof websites.

Web spams are involved in bypassing search engines to increase their rankings and earn more money [18]. It is further of two types i.e. boosting and hiding. Boosting involves intentional and false increase of value of web page using term, URL or link spamming. In hiding, the boosting and other unethical activities are intended to be hidden by content hiding, cloaking and redirection [19]. The second type of fake website is concocted website. Such websites appear to be genuine. For example, a fake concocted website of services or products provider where the payments are received from the customers but the customers do not get the services or products [20]. Spoof websites take advantage of information of customers because they are replica of authentic marketable websites and customers are redirected to the spoof websites where they become prone to frauds [20].

Many websites create blogs where unauthentic news and articles are published to enhance traffic and earn money. This gives rise to fake news production and spread [15]. Fake news refers to the false and manipulated information created and shared to mislead and deceive the audiences for financial and personal gain or causing harms [7]. Websites try to engage maximum number of audience and in doing so; they create fake news and articles with false information so that the users might be redirected to their websites when searching for specific terms on internet [5]. Political polarizations, easy access to revenues from online advertisements and high prevalence of social media have been implicated in the spread of fake news. Fake news types range from misleading content, parody, false connection and context, towards manipulated and fabricated content [21]. Therefore, fake news detection and tackling is crucial and challenging.

Nejad et al. [22] argued that sentiment analysis is practical approach in detection of spam and fake news. Sentiment analysis entails natural language processing and text analysis for affective states and subjective pieces of information. This is also referred as opinion mining from the datasets. Sentiment analysis is broadly classified into three categories i.e. lexicon based approach, machine learning approach and hybrid approach [23]. In lexicon based approach, unsupervised learning method is employed for data mining and it does not require previous training. The sentiments of words are polarized as negative or positive opinions. Lexicon based approach entails dictionary based and corpus based approach which further entails statistical and semantic approach [23]. In contrast, machine learning approach of sentiment analysis uses classification algorithms that classify the sentiments based on the trained labelled corpus for identifying features [24]. It can be supervised, unsupervised or semi-supervised. Unsupervised algorithms and clustering techniques have been widely used in literature including K-mediod, K-mean, neural networks and Gaussian mixtures. In terms of supervised algorithms, Naïve Bayes, maximum entropy and support vector machine (SVM) have been used in literature for classification of datasets [24]. Hybrid approach combines both lexicon and machine learning approaches of sentiment analysis.

Researchers have used all of these approaches for the classification and clustering of information from the datasets in the literature. The problem is with the accuracy of the developed models to detect fake news. For example, Ubing et al. [25] used hybrid approach and some algorithms like logistics regression, random forest and prediction model to assess fake websites and got 70% to 92% accuracy. On the other hand, Vicario et al. [26] used logistic regression, K-nearest, SVM, decision tree and neural networks to detect fake news topics and polarize contents on social media and achieved accuracy of 91%. To detect the abusive language on English Wikipedia community, Rawat et al. [27] developed a model based on machine learning algorithm i.e. XGBoost classifier and achieved accuracy of 84%. Similarly, detecting fake reviews on electronic business websites, Barbado et al. [28] developed sentiment analysis model based on Ada Boost classifier and achieved accuracy of 82%. Bharadwaj et al. [29] developed a hybrid model of sentiment analysis using semantic unigram TF, TFIDF, random forest, RNN and Naïve Bayes and achieved accuracy of 95.66% for the detection of fake news articles. This analysis shows that accuracy is the most problematic aspect in spam detection using sentiment analysis because accuracy explains the reliability and validity of the proposed models. There is a need for the development of models of sentiment analysis with higher accuracy.

Zvarevashe et al. [30] conducted sentiment analysis for hotel customers' feedback and compared different classifiers including Naïve Bayes, Naïve Bayes multinomial, sequential minimal optimization and composite hypercube. Arif et al. [10] also conducted performance evaluation of learning based classifiers for spam SMS and email from datasets of social media evaluations. The authors used XCS, XCSR and XCSR# algorithms and found that XCSR# is much better for fake SMS and email detection. Similarly, to detect the spam SMS using machine learning classifiers, Gupta et al. [31] performed comparison of performances of different algorithms and found that CNN (Convolutional Neural Network) classifier is the most effective one with accuracy of 91%. Deokate [32] also performed fake news detection using machine learning algorithms on Twitter datasets and found that support vector machine is most suitable for Twitter fake news detection. Similarly, Mathur et al. [33] also performed sentiment analysis on Twitter articles for abusive and offensive tweets. They found that convolutional neural network i.e. Ternary Trans-CNN is the most effective method for offensive tweets detection using sentiment analysis.

The literature review demonstrates that spam detection is an important and crucial field of research and is most challenging because of its complexity, wide number of available options to carry it out, wide number of sources of spam i.e. fake news, emails, SMS, tweets etc. and wide number of fake websites and blogs that publish spam. The previous studies show that most of the work is carried out for emails, SMS and tweets and little work is carried out for fake news. The researchers used different methods with different accuracies on random datasets without web scrapping and extraction of useful information. There is a research gap for fake news detection using web scrapper, hybrid sentiment analysis and validation of accuracy for the developed hybrid sentiment analysis approach for fake news detection. This research is carried out in this direction and addresses this research gap.

3 Methodology

The methodology entails the development of a model using hybrid approach of sentiment analysis i.e. machine learning and lexicon approaches combined. The model also entails a web scrapper tool to collect and analyze the data. The news related spam websites are

targeted and Kaggle datasets are also included in the collected data. The web scrapper tool is developed to scrape the contents of the news websites. The results are then generated by classification algorithms of machine learning and lexicon based approaches. The accuracy is also presented for these algorithms and they can classify the websites as spam or ham. The selected methodology of sentiment analysis is hybrid approach. The individual news is treated as individual entity. The proposed methodology is shown in the image below:

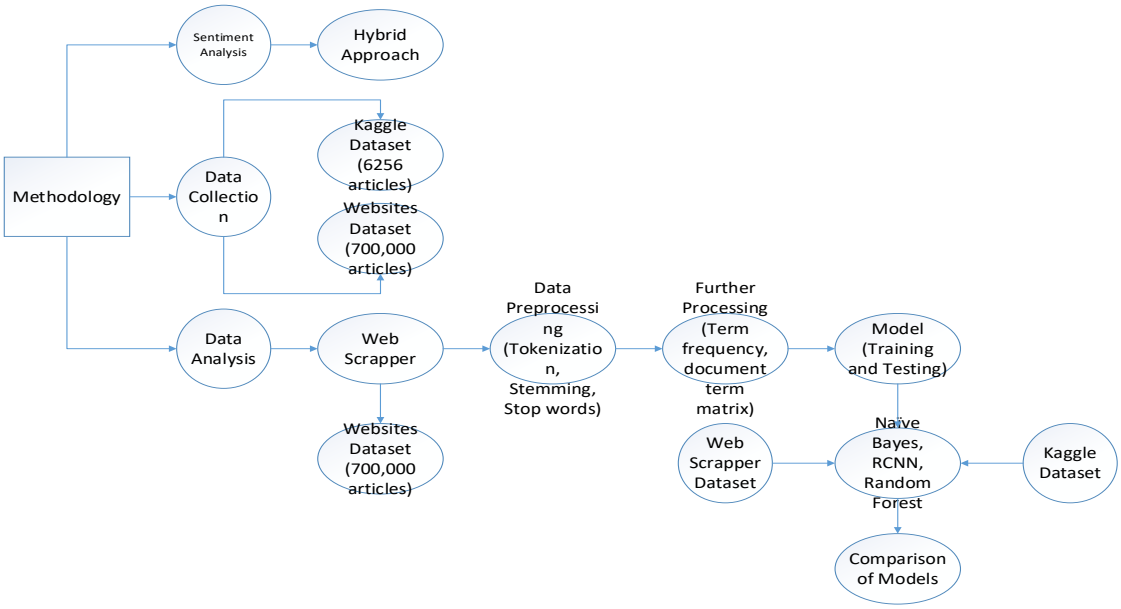

Figure 1 Proposed Methodology

3.1. Sentiment Analysis Approach

Hybrid sentiment analysis approach is selected for this research to make the results more reliable and valid and also analyse the data more vigorously. Hybrid sentiment analysis approach involves both machine learning approach and lexicon approach. In terms of machine learning approach, different classification algorithms are used which include Naïve Bays, Random Forest and RCNN. RCNN is used because it is under LSTM and has been rarely used before.

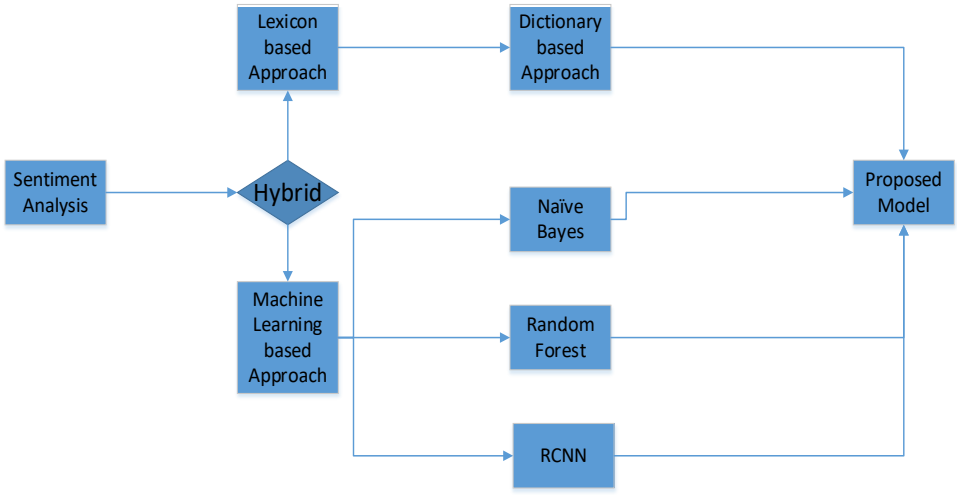

Figure 2 Proposed Sentiment Analyses

In terms of lexicon approach, dictionary based approach is used to classify the data extracted using web scrapper tool. Different Python libraries are used which include beautiful soup, numpy, pandas and request.

3.2. Data Collection

The data is collected from two sources. First, the news articles from different websites are collected and run through web scrapper tool

developed. The selected websites include news websites and approximately 700,000 news articles. Second source of data collection is Kaggle dataset. This data is collected from [34]. This dataset includes 6256 news articles as available on Kaggle too.

3.3. Data Analysis

The data analysis phase consists of various steps. These are as follows:

1. Data extraction using web scrapper.
2. Data cleaning and preparation.
3. Training and testing different models.
4. Performance measurement
5. Comparison of models for most accurate results.

3.3.1. Data Extraction via Web Scrapper

This phase involves gathering of useful information from unstructured text obtained from news articles. For this purpose, a web scrapper is developed with defined attributes and characteristics that help in extracting data form the websites. Beautiful soup, request, numpy and pandas are used to develop web scrapper. The scrapped information is exported to MS Excel file. The following list shows the data that has to be scrapped from the websites:

1. **Domain's Age:** It checks the time period of the domain name. If a registered time of the website is for less than 12 months, it is regard as a fake website.
2. **Known Logos:** It helps in checking if a web-page contains well-known logos that are not their real domains, which indicates a fake website.
3. **Wary URL:** it checks whether a webpage's URL contains an (@) 'at the rate 'sign or (-) a dash sign in there domain name, which likely exists in a fake website.
4. **Wary Links**: A fake website contains a broken links. It checks if there are any broken links.
5. **IP Name/Address:** It checks a domain name which is used instead of an IP address.
6. **No-of-Dots in URL**: Authentic URL contains a few numbers of dots. So, it helps in checking the dots in a URL. There are many dots in fake webpage's URL.
7. **HTML Forms**: HTML forms ask information that is personal/sensitive. It helps in checking for an HTML text entry form in a webpage.
8. **Images**: Use of images from existing legitimate or fake sites is so frequent. The images of products, employees, customers and company belongings are reused.
9. **Page Text:** In fake websites, the text of the web page has more likely mistakes in spelling and grammar.
10. **False Content:** The fake sites have the misleading contents. They produce the information that misleads the audience or visitors.
11. **False Connections**: The images which are given in the post having captions that do not give any accurate information about that image. There is no true meaning and relation to that text description given below of that image.

These attributes of web scrapper are used to extract the information from the collected data from news websites. Kaggle dataset is not run through web scrapper.

3.3.2. Data Cleaning and Preparation

This phase further cleans and prepares the data that is run through web scrapper. It has the following steps:

1. Tokenization: It is the process of fragmenting data or text into words, phrase or sentence, separate the blank-space, special characters e.g. punctuation marks, nouns, verbs, adverbs, adjectives, URLs etc.
2. Removal of Stop Words: The model is programmed to remove some words which are listed as stop word in manually created list of stop words. These are removed in order to save time and space.
3. Stemming: By the studying of structure of words and parts of words, stemming is done. It is a process of removing the affixes, plural genders and conjugation.

After this phase, term frequency is computed and a document term matrix or term-document matrix is developed. It describes the frequency of terms that occur in a collection of documents, in the form of rows and columns.

3.3.3. Training and Testing Models

In this phase, determination of values to the training matrixes is carried out and the models generate output values based on the inputs. The input values in this phase entail fake news articles and websites that have been scrapped through web scrapper. The models used in this step are machine learning supervised classifiers and include Naïve Bayes, RCNN (Neural Network) and Random Forest. These models are the best choice for classification and are very effective for spam detection and classification. The Kaggle dataset is also run through these models. The models are trained and tested and outputs are generated in this phase for both types of datasets i.e. web scrapper dataset and Kaggle dataset.

3.3.4. Performance Measurement

The performance of the proposed model along with the supervised machine learning algorithms is measured in this phase. This is carried out with the help of confusion matrices. The outputs from the models are evaluated and confusion matrices are formed for each algorithm and for

each dataset. Confusion matrix explains the performance of a machine learning algorithm with two or more classes as outputs. It is presented as a matrix with actual and predicted values.

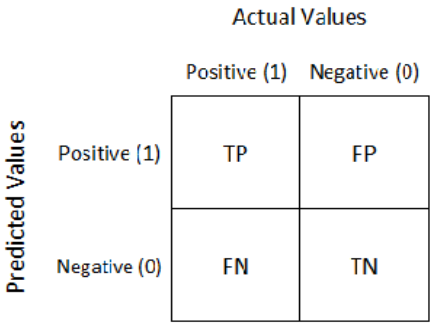

Figure 3 Confusion Matrix

In this aspect, the four values i.e. true positive, false negative, true negative and false positive are used as the inputs for measuring accuracy, precision, recall and F-values. These measures are used to evaluate performance of the proposed model and algorithms. Following equations are used to measure them:

$$Accuracy = TP + TN \div TP + TN + FP + FN$$

$$Recall = TP \div TP + FN$$

$$Precision = TP \div TP + FP$$

$$F_{measure} = 2 \times Recall \times Precision \div Recall + Precision$$

3.3.5. Comparison of Models

The models are then compared for both types of datasets. The results are compared to analyze which model is best suited for spam detection including the output values for both of the datasets i.e. web scrapper dataset and Kaggle dataset.

4 Results and Discussions

The results are presented and discussed in this section along with the comparisons of different algorithms. The results for web scrapper dataset and Kaggle dataset are presented individually.

4.1 Kaggle Dataset

All of the algorithms i.e. Naïve Bayes, RCNN and Random Forest were implemented for Kaggle dataset. The model developed test sets for the dataset. These test tests were analysed by the algorithms. The models were trained for the classification of fake news. The results from these models are presented in confusion matrix form. Each matrix contains true labels and predicted labels for the credible and non-credible news i.e. real and fake.

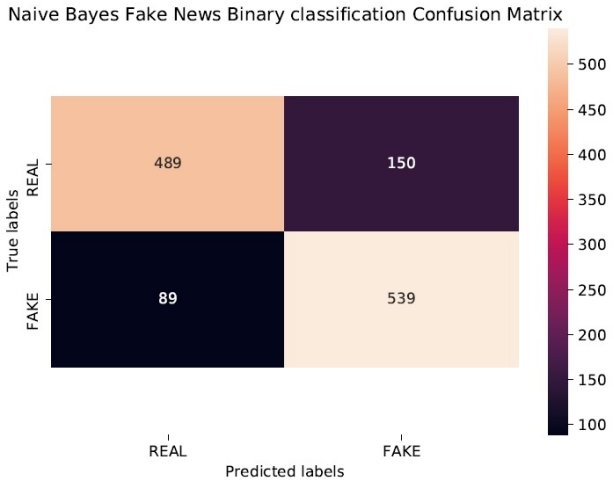

Figure 4 Naive Bayes Confusion Matrix for Kaggle Dataset

The test sets developed by the model (Naïve Bayes) include 489 true positive news articles that mean that these news articles are observed positive and also predicted to be positive i.e. real. 89 news articles are false positive i.e. they are observed as fake but these articles are predicted as real. 539 articles are observed to be fake and also predicted to be fake i.e. true negative. 150 articles are observed as real but are predicted as fake i.e. false negative.

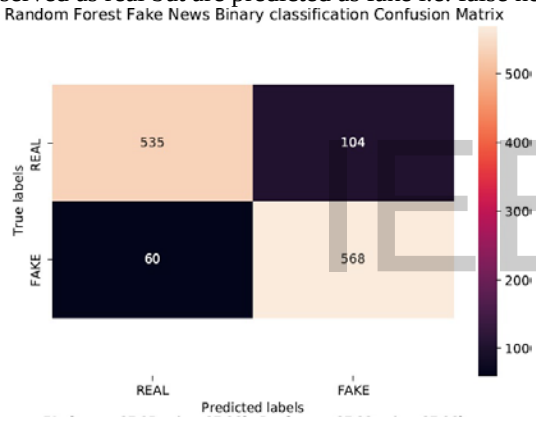

Figure 5 Random Forest Confusion Matrix for Kaggle Dataset

In Random Forest confusion matrix, 535 articles are predicted to be real and observed real too. 60 articles are observed as fake but predicted as real. 568 articles are observed and predicted as fake and 104 articles are observed as real but are predicted as fake.

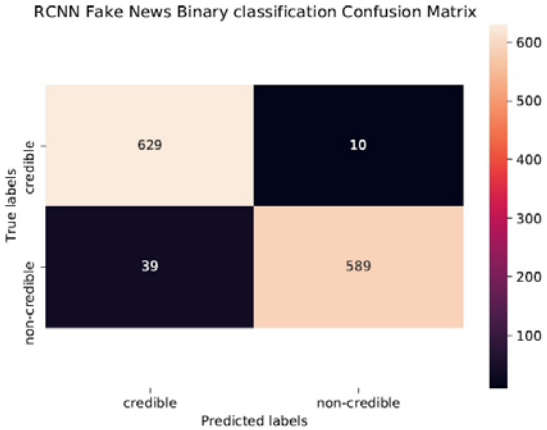

Figure 6 RCNN Confusion Matrix for Kaggle Dataset

In RCNN model, 629 articles are real as predicted and observed. 39 articles seem as fake but are predicted as real. 589 articles are

considered as fake entirely because the observations and predictions deem them fake. 10 articles are observed as real but are predicted as fake.
The precision, recall, accuracy and f-measures for Kaggle dataset are given in the table below:

Table 1 Performance Measures

Model	**Performance Measures**			
Name	**Precision [P]**	**Recall [R]**	**F-Measure**	**Accuracy**
Naïve Bayes	81.42	81.18	81.11	81.14
Random Forest	87.22	87.09	87.05	87.06
RCNN	96.25	96.11	96.13	96.13

These values show that the least effective model is Naïve Bayes and the most effective model is RCNN. This means that the spam detection using Kaggle dataset and RCNN generates highest accuracy of 96.13. Other models have also generated high accuracies but RCNN is the most accurate and valid model for spam detection using hybrid sentiment analysis.

4.2 Web Scrapper Dataset

As with Kaggle dataset, web scrapper dataset was also run through all of the models i.e. Naïve Bayes, Random Forest and RCNN. The dataset consisted of 700,000 news articles but the test sets were developed by each model differently including or discarding different number of news articles. The confusion matrices for these models and dataset are presented in the following figures.

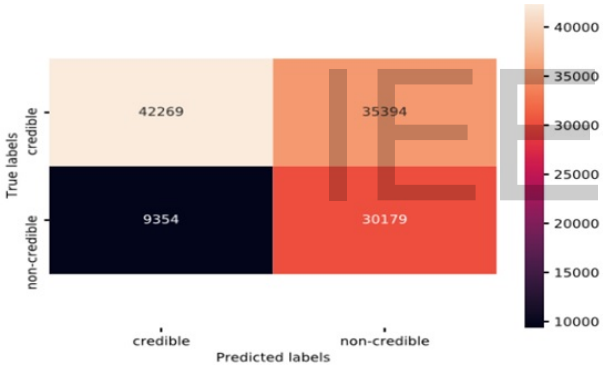

Figure 7 Nave Bayes Confusion Matrix for Web Scrapper Dataset

Nave Bayes model shows that 422269 articles are credible and real and 30179 articles are entirely considered as fake ones as observed and predicted. 9354 articles are observed to be fake ones but are predicted as credible and real. 35394 articles are observed as real and credible but predicted as fake ones.

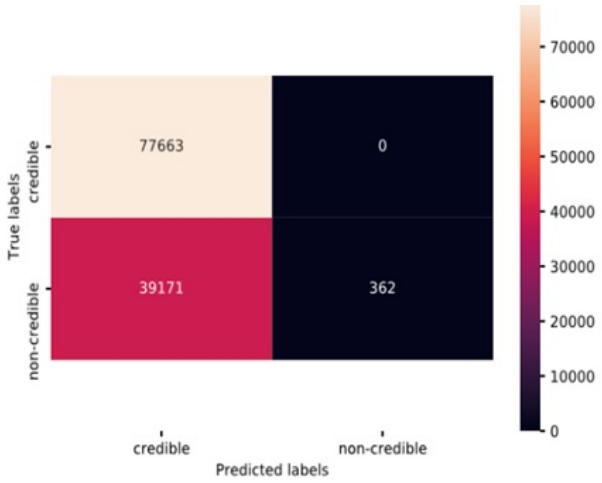

Figure 8 Random Forest Confusion Matrix for Web Scrapper Dataset

In Random Forest model, 77663 articles are credible and real while 362 articles are fake ones as observed and predicted. 39171 articles are observed as fake ones but are real and credible as predicted.
RCNN model shows that 70952 news articles are real because they are observed and predicted to be real. 9880 news articles are observed as fake ones but are real and credible as predicted. However, 29653 articles are considered as fake as observed and predicted. 6711 news articles are observed as real but are fake ones as predicted.

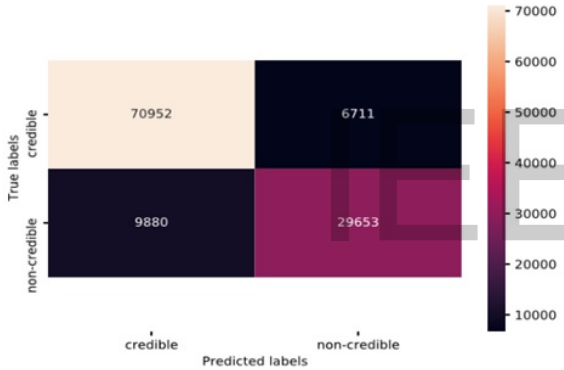

Figure 9 RCNN Confusion Matrix for Web Scrapper Dataset

The performance measures for web scrapper dataset for the developed models are shown in the following table.

Table 2 Performance Measures for Web Scrapper Dataset

Model	**Performance Measures**			
Name	**Precision [P]**	**Recall [R]**	**F-Measure**	**Accuracy**
Naïve Bayes	63.95	65.38	61.41	61.82
Random Forest	83.24	50.46	40.84	66.58
RCNN	85.23	84.22	84.68	86.5

The table shows that the least effective and accurate method in case of web scrapper dataset is again Naïve Bayes. The most efficient and accurate model is again RCNN in case of web scrapper dataset as that of Kaggle dataset.

4.3 Comparison of Models

The comparison of all of the models and algorithms for both datasets is presented in the following table.

Model Name	Accuracy	
	Kaggle Dataset	Web Scrapper Dataset
Naïve Bayes	81.14	61.82
Random Forest	87.06	66.58
RCNN	96.13	86.5

The table shows that using the web scrapper data, the highest accuracy came out to be 86.5. This was considered as relatively low because of the prevalence of achievements of higher accuracies in the recent studies. Therefore, Kaggle dataset was also employed by the researcher to validate the developed models and check their accuracy and reliability. The main purpose of these models was to detect the spam i.e. fake news from the given datasets. The news articles were treated as individual entities in both of the datasets and all of the models. The accuracies for the Kaggle dataset were found to be higher for all of the developed models as compared with web scrapper dataset. Although the accuracies were relatively lower for web scrapper data as compared with that of Kaggle, yet the spam detection was carried out successfully by the developed models. The highest accuracy in case of web scrapper data was 86.5 and in case of Kaggle data was 96.13. These values indicate that the developed models are valid and accurate and can successfully detect spam and fake news from the datasets. In these models, RCNN is the best suited model for spam detection and fake news or websites detection because it has the highest accuracy in both of the datasets. Naïve Bayes method is also practical but less accurate. Another option is Random Forest model for spam and fake news detection but it is also less accurate as compared with RCNN. Therefore, the results indicate that spam detection using hybrid sentiment analysis approach is best carried out using RCNN algorithm.

5 Conclusion and Future work

Millions of fake news are generated and spread across internet by fake websites daily. These types of fake news are targeted for suspicious, unethical and criminal activities by the perpetrators. The victims are from every domain i.e. individual internet users, financial institutions, and government websites, search engines and file hosting servers and websites. Spam detection is a highly complicated and challenging task and requires technical and advanced methodologies to be tackled. Sentiment analysis is the suitable answer for spam and fake news and websites detection. Therefore, the development of an effective methodology using hybrid sentiment analysis approach proved to be effective for fake news detection in this research. The lexicon based and machine learning based sentiment analysis approach utilized different models and algorithms for spam and fake news detection. These algorithms included Naïve Bayes, Random Forest and RCNN. Different websites were used for data collection and also a dataset from Kaggle was also utilized. Web scrapper tool was developed to extract information from the raw and unstructured collected data. The collected data from both sources was run through the developed models and the fake news articles were identified. The models proved to be effective in detecting spam and fake news. Different performance measures were developed and performance of each model for both datasets was measured. The most efficient and accurate model was RCNN model for spam and fake news detection. Using this model, fake news can be identified from the internet and other datasets and the results can be used by the regulatory authorities to ban such websites that spread fake news. This would be beneficial for everyone because removing fake news and spam from internet will make this era more reliable and safe. The future direction can be implementation of the proposed model for SMS spam, email spam and social media spam detection and addition of more classification algorithms to enhance the accuracy and validity of the model.

6 Conflicts of interest

The authors have no conflict of interest to declare.

7 References

1. Paul A, Jeyaraj R. Internet of Things: A primer. Human Behavior and Emerging Technologies. 2019 Jan;1(1):37-47.
2. Atzori L, Iera A, Morabito G. Understanding the Internet of Things: definition, potentials, and societal role of a fast evolving paradigm. Ad Hoc Networks. 2017 Mar 1;56:122-40.
3. MacDermott Á, Baker T, Buck P, Iqbal F, Shi Q. The Internet of Things: Challenges and considerations for cybercrime investigations and digital forensics. International Journal of Digital Crime and Forensics (IJDCF). 2020 Jan 1;12(1):1-3.
4. Sarmah A, Sarmah R, Baruah AJ. A brief study on Cyber Crime and Cyber Law's of India. International Research Journal of Engineering and Technology (IRJET). 2017 Jun;4(6):1633-40.
5. Zhou X, Zafarani R, Shu K, Liu H. Fake news: Fundamental theories, detection strategies and challenges. InProceedings of the twelfth ACM international conference on web search and data mining 2019 Jan 30 (pp. 836-837).
6. Ahmed H, Traore I, Saad S. Detecting opinion spams and fake news using text classification. Security and Privacy. 2018 Jan;1(1):e9.
7. Mustafaraj E, Metaxas PT. The fake news spreading plague: was it preventable?. InProceedings of the 2017 ACM on web science conference 2017 Jun 25 (pp. 235-239).

8. Jelodar H, Wang Y, Yuan C, Jiang X. A systematic framework to discover pattern for web spam classification. In2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) 2017 Oct 3 (pp. 32-39). IEEE.
9. Saumya S, Singh JP. Detection of spam reviews: a sentiment analysis approach. Csi Transactions on ICT. 2018 Jun;6(2):137-48.
10. Arif MH, Li J, Iqbal M, Liu K. Sentiment analysis and spam detection in short informal text using learning classifier systems. Soft Computing. 2018 Nov;22(21):7281-91.
11. Asghar MZ, Ullah A, Ahmad S, Khan A. Opinion spam detection framework using hybrid classification scheme. Soft computing. 2020 Mar;24(5):3475-98.
12. Mahajan S, Rana V. Spam detection on social network through sentiment analysis. Advances in Computational Sciences and Technology. 2017;10(8):2225-31.
13. Kaggle. Kaggle: Your Machine Learning and Data Science Community [Internet]. Kaggle.com. 2021 [cited 25 February 2021]. Available from: https://www.kaggle.com/
14. Agogo D, Hess TJ. "How does tech make you feel?" a review and examination of negative affective responses to technology use. European Journal of Information Systems. 2018 Sep 3;27(5):570-99.
15. Quandt T, Frischlich L, Boberg S, Schatto-Eckrodt T. Fake news. The international encyclopedia of Journalism Studies. 2019 May 14:1-6.
16. Watson A. Frequency of fake news on online news websites U.S. 2020 | Statista [Internet]. Statista. 2020 [cited 25 February 2021]. Available from: https://www.statista.com/statistics/649234/fake-news-exposure-usa/
17. Carpineto C, Romano G. Learning to detect and measure fake ecommerce websites in search-engine results. InProceedings of the international conference on web intelligence 2017 Aug 23 (pp. 403-410).
18. Park AJ, Quadari RN, Tsang HH. Phishing website detection framework through web scraping and data mining. In2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) 2017 Oct 3 (pp. 680-684). IEEE.
19. Makkar A, Kumar N. An efficient deep learning-based scheme for web spam detection in IoT environment. Future Generation Computer Systems. 2020 Jul 1;108:467-87.
20. Elnagar S, Thomas M. A cognitive framework for detecting phishing websites. InInternational Conference on Advances on Applied Cognitive Computing (ACC 2018) 2018 (pp. 60-61).
21. Sukhodolov AP, Bychkova AM. Fake news as a modern media phenomenon: definition, types, role of fake news and ways of counteracting it. Вопросы теории и практики журналистики. 2017;6(2).
22. Nejad SJ, Ahmadi-Abkenari F, Bayat P. Opinion Spam Detection based on Supervised Sentiment Analysis Approach. In2020 10th International Conference on Computer and Knowledge Engineering (ICCKE) 2020 Oct 29 (pp. 209-214). IEEE.
23. Yusof NN, Mohamed A, Abdul-Rahman S. Reviewing classification approaches in sentiment analysis. InInternational conference on soft computing in data science 2015 Sep 2 (pp. 43-53). Springer, Singapore.
24. Agarwal B, Mittal N. Machine learning approach for sentiment analysis. InProminent feature extraction for sentiment analysis 2016 (pp. 21-45). Springer, Cham.
25. Ubing AA, Jasmi SK, Abdullah A, Jhanjhi NZ, Supramaniam M. Phishing website detection: An improved accuracy through feature selection and ensemble learning. International Journal of Advanced Computer Science and Applications (IJACSA). 2019 Jan 1;10(1).
26. Vicario MD, Quattrociocchi W, Scala A, Zollo F. Polarization and fake news: Early warning of potential misinformation targets. ACM Transactions on the Web (TWEB). 2019 Mar 27;13(2):1-22.
27. Rawat C, Sarkar A, Singh S, Alvarado R, Rasberry L. Automatic Detection of Online Abuse and Analysis of Problematic Users in Wikipedia. In2019 Systems and Information Engineering Design Symposium (SIEDS) 2019 Apr 26 (pp. 1-6). IEEE.
28. Barbado R, Araque O, Iglesias CA. A framework for fake review detection in online consumer electronics retailers. Information Processing & Management. 2019 Jul 1;56(4):1234-44.
29. Bharadwaj P, Shao Z. Fake news detection with semantic features and text mining. International Journal on Natural Language Computing (IJNLC) Vol. 2019 Jul 24;8.
30. Zvarevashe K, Olugbara OO. A framework for sentiment analysis with opinion mining of hotel reviews. In2018 Conference on information communications technology and society (ICTAS) 2018 Mar 8 (pp. 1-4). IEEE.
31. Gupta M, Bakliwal A, Agarwal S, Mehndiratta P. A comparative study of spam SMS detection using machine learning classifiers. In2018 Eleventh International Conference on Contemporary Computing (IC3) 2018 Aug 2 (pp. 1-7). IEEE.
32. Suchitra B. Deokate," Fake News Detection using Support Vector Machine learning Algorithm", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177 Volume 7 Issue VII, July 2019
33. Mathur P, Shah R, Sawhney R, Mahata D. Detecting offensive tweets in hindi-english code-switched language. InProceedings of the Sixth International Workshop on Natural Language Processing for Social Media 2018 Jul (pp. 18-26).
34. Nørregaard J, Horne BD, Adalı S. NELA-GT-2018: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles. ICWSM [Internet]. 2019Jul.6 [cited 2021Feb.25];13(01):630-8. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/3261