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.

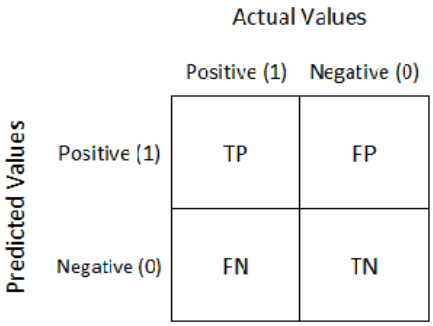

Figure 3 Confusion Matrix

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.

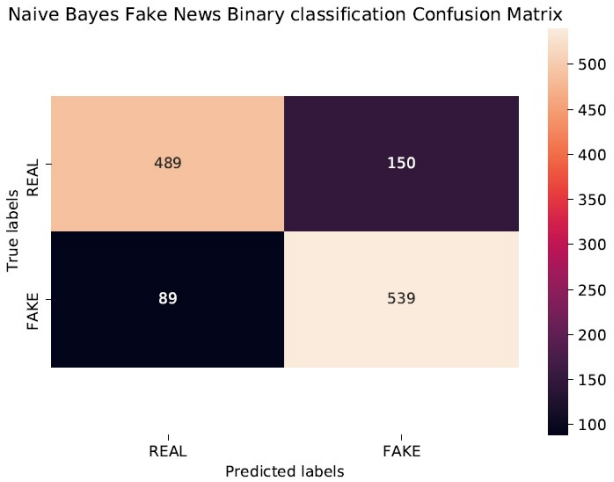

Figure 4 Naive Bayes Confusion Matrix for Kaggle Dataset

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.

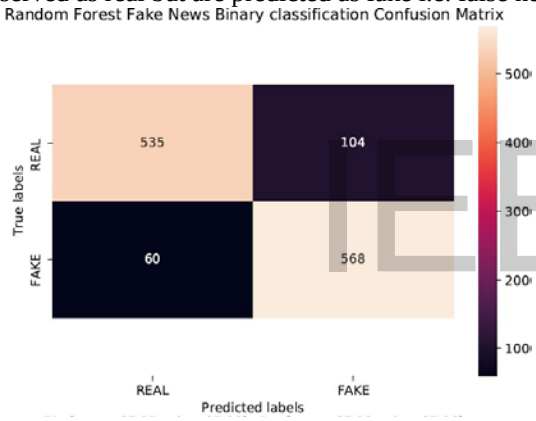

Figure 5 Random Forest Confusion Matrix for Kaggle Dataset

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.

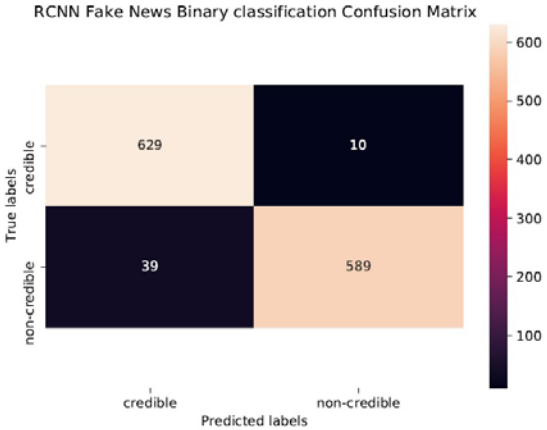

Figure 6 RCNN Confusion Matrix for Kaggle Dataset

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.

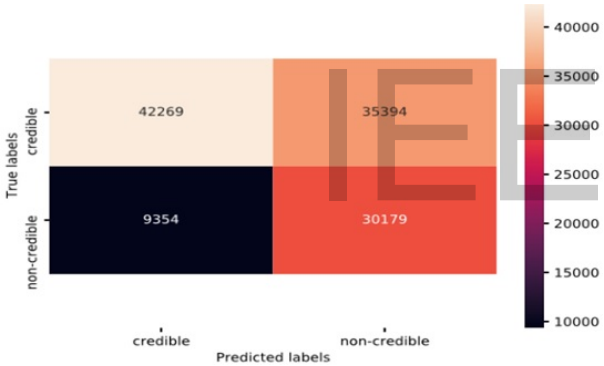

Figure 7 Nave Bayes Confusion Matrix for Web Scrapper Dataset

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.

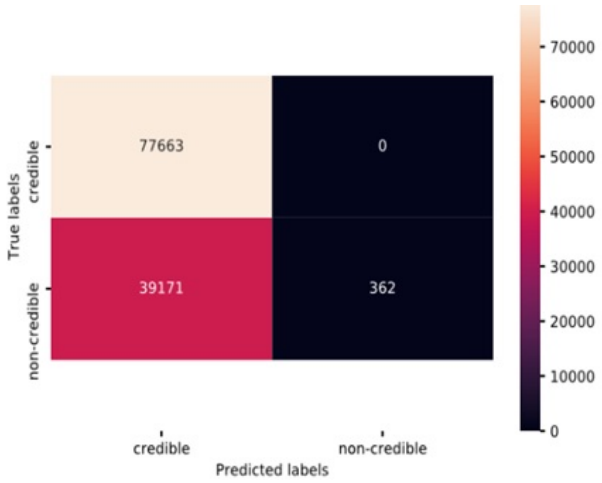

Figure 8 Random Forest Confusion Matrix for Web Scrapper Dataset

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.

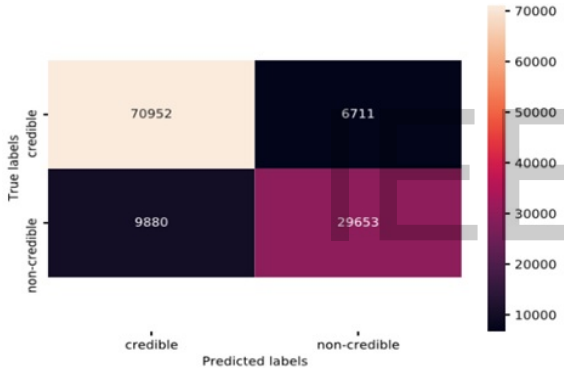

Figure 9 RCNN Confusion Matrix for Web Scrapper Dataset

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