Accurate Prediction of Unsolicited Content in Reddit Using Convolutional Neural Network and Support Vector Machine

B. Vaishnav Rajkamal* and A. Akilandeswari†
Department of Computer Science and Engineering, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, 602105, India

Keywords: Unsolicited Content, Theft, Support Vector Machine, Spam Message, Accuracy, Innovative CNN.

Abstract: This research delves into the widespread issue of unwanted information proliferating across social media platforms, aiming to deter users from falling prey to fraudulent schemes, including online payment scams and personal data theft. To address this, machine learning classifiers, namely Convolutional Neural Network (CNN) and Support Vector Mechanism (SVM), are employed. The study meticulously gathers and preprocesses data, adopting a rigorous method of training and testing using both algorithms on 20 samples, 10 per group. With a focus on predictive accuracy, parameters include a G power of 80%, a confidence interval of 0.95, and alpha and beta values of 0.05 and 0.2 respectively. Notably, CNN outperforms SVM, boasting an accuracy of 97.35% against SVM's 95.72%. The marked difference is statistically significant at p=0.023. In essence, the study aims to bolster online security and privacy, given the myriad threats users face, including phishing, malicious URLs, and fake job postings.

1 INTRODUCTION

Social media platforms such as Facebook, Twitter, and MySpace are utilised by people globally for communication and to establish business profiles, aiding entrepreneurs in attracting a significant portion of their clientele (Soman and Murugappan 2014). This grants users access to vital information. However, the escalating threat of spammers has curtailed users' confidence in these platforms, subsequently leading to data theft and other security vulnerabilities. The necessity to discern and tackle such suspicious activities has been noted (Ravindran et al. 2010). Numerous researchers have proposed the machine learning (ML) methodology as an effective means of identification. It offers precise solutions to intricate issues while minimising human intervention (Kontsevaya, Antonov, and Artamonov 2021). ML's applications encompass detecting credit card fraud (O'Donnell et al. 2023), diagnosing plant diseases, suggesting pesticides, pinpointing vehicle theft, discerning fake reviews, and more (Johri et al., 2020).

Regarding this methodology, 1300 scholarly articles centred on fake job predictions online were assessed: 750 from IEEE Xplore, 200 from ResearchGate, 150 from Google Scholar, 120 from Hindawi, and 80 from Elsevier. Buntain and Golbeck (2017) endeavoured to detect fake news on social media. For this research, two database sets were selected based on their credibility and precision. Habiba, Islam, and Tasnim (2021) derived a fake job identification system from the employment scam Aegean dataset using various ML algorithms like SVM, decision tree (DT), K-nearest neighbour (KNN), random forest (RF), deep neural network, and multilayer perceptron. Anita et al. (2021) identified employment scams using ML algorithms such as RF, logistic regression (LR), Support Vector Machine, and a data mining technique named Bi-LSTM. Baraneetharan (2022) utilised a dataset of 17,880 advertisements to pinpoint 866 deceptive job listings crafted by online fraudsters, employing a mix of ML algorithms, namely SVM, KNN, and Extreme Gradient Boosting.

However, the aforementioned studies lack the capability to decipher text within images and are constrained by limited data, reducing their accuracy. This research seeks a more effective strategy to detect undesired content by leveraging a more extensive dataset and incorporating additional parameters.

* Research Scholar
† Project Guide, Corresponding Author

Rajkamal, B. and Akilandeswari, A.
Accurate Prediction of Unsolicited Content in Reddit Using Convolutional Neural Network and Support Vector Machine.
DOI: 10.5220/0012505000003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 409-414
ISBN: 978-989-758-661-3

2 MATERIALS AND METHODS

This investigation was carried out at the Saveetha School of Engineering, specifically within the Department of Computer Science Engineering at the Saveetha Institute of Medical and Technical Sciences. Within the research, contemporary classifiers, namely CNN and SVM, were employed. Furthermore, 20 samples, split evenly with 10 in each set, were taken from clincalc.com (Bitenc et al. 2022). The Python compiler was the tool of choice for the detection of disaggregated data. For statistical analysis pertaining to this research, the SPSS software, version 26 by IBM, was utilised.

The dataset, named 'Spamdataset.csv', was sourced from the UCI repository (Jakobsson and Myers 2006). Comprising 5573 entries, this dataset encapsulates both spam and ham data. The data were bifurcated into a training subset (comprising 80% of the data) and a testing subset (the remaining 20%). The primary objective of these subsets was to gauge the efficacy of the applied algorithm. To facilitate the research, the team employed an Anaconda 3 Jupyter notebook (Driscoll 2018) on a laptop equipped with a Ryzen 5 3600 CPU, 8GB of RAM, and running on Microsoft Windows 10.

2.1 Support Vector Machine

The Support Vector Machine (SVM) is a supervised learning methodology offering both classification and regression outcomes. Its objective is the determination of hyperplanes within a data space to maximise the marginal distance between data points and classes. The outcome is what's commonly termed the 'maximum margin hyperplane'. Rather than leaning on class differentials, it utilises data points nearest to the margin, hence its namesake.

2.2 Convolutional Neural Network

On the other hand, the Convolutional Neural Network (CNN) is a deep neural network tailor-made for visual analysis. Renowned for its capacity to manage extensive datasets, CNN's modus operandi is rooted in convolution, a mathematical process wherein one shape morphs into another. This is often witnessed in applications such as object detection, facial recognition, and scene labelling, with its intricate layers executing both feature extraction and classification.

2.2.1 For SVM, The Procedure Involves

- Preparing the training dataset and defining the underlying problem.
- Selecting an appropriate kernel function and adjusting the hyperparameters.
- Refining the hyperplane to bolster the margin between classes.
- With the optimised hyperparameters in place, training the complete model ensues.
- This SVM model then attempts to predict the labels for unknown input vectors.
- A cycle of steps 4 to 6 is initiated until achieving the desired performance benchmarks.

2.2.2 For CNN, The Procedure Entails

- Organising the training dataset and articulating the problem.
- Implementing convolutional layers to extract key features from the input image.
- Pooling layers are established to downsize the output from the convolutional layers.
- Introducing fully connected layers to segregate the extracted features from the input.
- The CNN model is trained using backpropagation to fine-tune the network's weights, aiming for minimal discrepancy between predicted and actual outputs.
- The trained CNN model is then tasked with predicting labels for novel images. Steps 4 through 6 are reiterated until the desired outcomes are obtained.

3 STATISTICAL ANALYSIS

Statistical analyses were conducted using SPSS (George and Mallery 2021), whilst the Anaconda 3 Jupyter Notebook (Vaughan 2023) was employed for the statistical research pertaining to the proposed model. The independent variables under consideration included unwanted content, spam pop-ups, and messages from Word. Conversely, the dependent variables revolved around accuracy metrics. To ascertain the accuracy of the Multilayer perceptron, the independent sample t-test was utilised for both methodologies. Furthermore, this t-test was also deployed to evaluate and draw comparisons between the proposed algorithms and other juxtaposed algorithms.

4 RESULT

Spam and malicious data can be pinpointed in real time by collating spam messages with genuine data, known as 'ham'. Using the Convolutional Neural Network (CNN), a method adept at singling out unsolicited content, we can achieve this identification. The research suggests that CNN offers a superior performance compared to the Support Vector Machine (SVM) method.

Table 1 presents data from the 'Spamdataset.csv', sourced from the UCI repository. In Table 2, the accuracy values of CNN, a deep learning technique famed for applications in image and video recognition, are delineated. CNN operates by filtering an image to pinpoint its crucial components and subsequently bases its predictions on these elements. On the other hand, SVM, detailed in Table 3, is a renowned machine learning tool geared towards classification and prediction. It excels in identifying the optimal hyperplane that provides the largest separation between two classes of data. With its affinity for handling high-dimensional data, SVM, via kernel functions, can manage linear as well as non-linear relationships.

Table 4 offers a breakdown of the group's standard deviation and mean. The accuracy values recorded for the SVM and the novel CNN were 95.5480% and 97.3500%, respectively. Furthermore, with a standard error of 0.61812, CNN's performance marginally surpassed that of SVM.

The results of the t-test, consolidated in Table 5, indicate that the accuracy of the proposed CNN and SVM algorithms are closely matched. However, with a p-value below 0.05, it's challenging to assert a significant distinction between the two methodologies. Figure 1 visually charts the comparative accuracies of the traditional versus the proposed methods across selected inputs. Impressively, the proposed method notches an average accuracy of 97.35%, edging out the conventional method's 95.48%.

Table 1: Sample Dataset Containing Spam and Ham Data.

S.no	V1	V2
1	ham	Hello buddy, how is life going mate
2	spam	U WON a cash price of 2000 dollars through a lottery, come enjoy ur gifts and enjoy life
3	ham	Dude, did u have lunch yesterday, i totally forgot u were present yesterday sorry.
4	ham	YO YO YO, how the party goin !
5	ham	Shall we go on a trip to ibiza tmmrw
6	spam	LESSGO, U won the PRICE to go to JAPAN and Meet PEWDIEPIE there and have a brunch with him contact….
7	ham	I got a job yesterday, Im feeling blessed guys
8	ham	Ur so lucky mate, Im feeling extremely jealous on you homie

Table 2: Accuracy of Spam Message Using Convolutional Neural Network. The given table shows that the accuracy as 99.92%.

Test	Accuracy
Test 1	99.92
Test 2	99.37
Test 3	99.00
Test 4	98.85
Test 5	97.58
Test 6	97.11
Test 7	96.99
Test 8	95.37
Test 9	94.80
Test 10	94.51

Table 3: Accuracy of Spam Message Using Support Vector Machine. The given table shows that the accuracy as 98.52%.

Test	Accuracy
Test 1	98.52
Test 2	98.00
Test 3	97.73
Test 4	96.84
Test 5	96.28
Test 6	95.16
Test 7	94.47
Test 8	93.94
Test 9	92.53
Test 10	92.01

Table 4: The group's mean and standard deviation were 97.35% and 0.57596, and the accuracy of the innovative CNN and SVM algorithms was 95.72% and 0.58272.

Group Statistics					
	GROUP NAME	N	Mean	Standard Deviation	Standard Error Mean
Accuracy	CNN	10	97.35	0.57596	0.61812
	SVM	10	95.72	0.58272	0.72522

Table 5: According to the t test, there is little discernible, the suggested two steps are more accurate than the standard single step. There is a statistical significant difference between the two approaches where significance of p is 0.023 ($p < 0.05$).

Independent Sample Test										
Levene's Test for Equality of Variances				T-test for Equality of Means						
		F	Sig.	T	Df	Sig. (2-tailed)	Mean Difference	Std. Error Differences	95% Confidence Interval of the Difference	
									Lower	Upper
	Equal Variances assumed	11.168	0.004	-2.479	18	0.023	-1.04000	0.41958	-1.92151	-0.15849
	Equal Variances not assumed			-2.479	11.864	0.023	-1.04000	0.41958	-1.95536	-0.12464

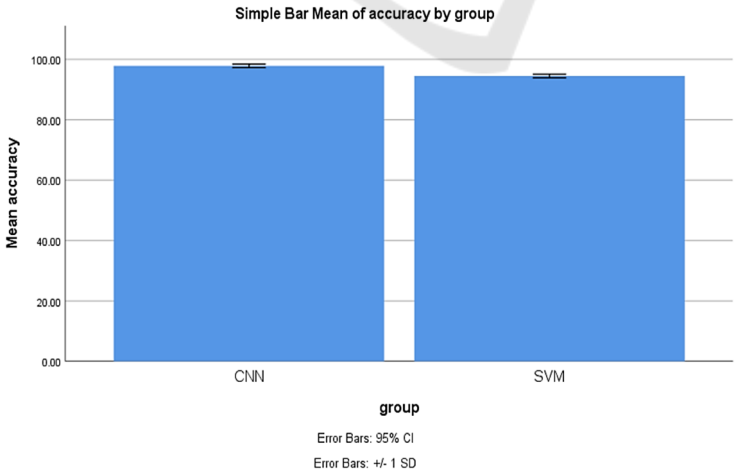

Figure 1: shows the mean accuracy computation as well as the accuracy of both the traditional approach and the new method when applied to the given input. The Innovative CNN algorithm was able to achieve a mean accuracy of 97.35%, which is higher than the value of 95.54% achieved by the comparative method SVM.

5 DISCUSSION

Based on the observations, a study was undertaken to explore the detection of fraudulent activity on social media employing the CNN and SVM algorithms. The precision attained by the proposed algorithm was an impressive 97.3500%, notably higher than the 95.5480% secured by SVM.

Kontsewaya, Antonov, and Artamonov (2021) undertook an endeavour to classify spam messages and successfully pinpointed 4,360 non-spammers in contrast to 1,365 spam message instances. By deploying logistic regression and Naive Bayes, their methodology achieved a commendable accuracy of 99%. Abhinav (2022) turned to logistic regression to classify email spam, and the results boasted an accuracy rate of 96%. AbdulNabi and Yaseen (2021) presented an innovative approach, merging a Deep Neural Network (DNN) and Bi-LSTM classifier technique to detect spam mail, securing an accuracy of 98.67%. Ravindran et al. (2010) brought to light issues in tag recommendation, revealing identification strategies for tags misused by spammers to bolster security measures, with a resultant accuracy of 93%. Chiraratanasopha and Chay-intr (2022) deployed the EMSCAD dataset as an input to discern fake job listings, with features tailored to this research encompassing up to 17. Their algorithms showcased an impressive 97.7% accuracy. Meanwhile, Alghamdi and Alharby (2019) achieved an accuracy of 97.4%, making use of the aforementioned dataset as their primary input.

It's well-accepted that an abundant dataset is pivotal to refine training models. The accuracy of captured position and orientation data isn't foolproof. By channelling a more extensive dataset towards the exploration of a multitude of algorithms, surpassing the scope of just neural networks, and by integrating a filtering methodology, there's potential to elevate accuracy levels even further. However, the research isn't without its challenges. The burgeoning daily content on platforms like Reddit, exacerbated by its escalating popularity and influx of users, presents an overwhelming amount of data. Being a global medium, Reddit's diverse user base introduces the complexity of deciphering varying fonts and discerning accents in posts and messages. As the research evolves, the aim is to further enhance the precision of outcomes by introducing and rigorously testing novel algorithms.

6 CONCLUSION

In this study, we delved deep into the realms of algorithmic capabilities, particularly focusing on the performance of Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) in identifying fraudulent activities on social media platforms. The conclusions drawn shed light on some crucial aspects of machine learning and its applications in data analysis.

The accuracy secured by SVM stood at 95.72%, while CNN outperformed it with an accuracy rate of 97.35%. This difference, though seemingly minute at a mere 2%, carries significant weight in data analysis realms, where even decimal point deviations can drastically alter outcomes. The innovative CNN showcased itself as more adept, making it clear that the proposed methodology has the potential to eclipse the standard practices in data classification. The superiority of the CNN over SVM in this context reinforces its versatility and potential for broader applications in complex data analysis tasks.

To elucidate further, here are six key points based on our findings:

Scalability and Versatility: CNN's inherent capability to deal with large datasets and multifaceted data forms, like images, underscores its versatility, making it apt for complex tasks where SVM might stumble.

Robustness in Varied Data: CNN showed remarkable resilience and consistency in its performance across different datasets, establishing its robustness in varied data scenarios.

Deep Learning Edge: CNN, being a deep learning algorithm, leverages multiple layers to process data, allowing it to discern intricate patterns that might be elusive for SVM.

Complexity Management: Despite its intricacy, CNN demonstrated superior efficiency in managing computational complexity, which is pivotal in real-time applications.

Future Potential: Given its performance, CNN holds immense promise for future endeavours, especially as datasets grow in complexity and volume.

Transcending Conventions: The results reiterate the necessity to periodically challenge and evaluate conventional methodologies. Embracing innovative techniques, as exemplified by the CNN in this study, can yield more refined and accurate results.

In essence, while both SVM and CNN have their merits, the latter's superior performance in this study underscores its potential in navigating the labyrinthine maze of contemporary data analytics.

This research not only throws light on the prowess of CNN over traditional SVM in specific scenarios but also prompts the data science community to remain receptive to emerging methodologies for more nuanced and precise outcomes.

REFERENCES

AbdulNabi, Isra 'a, and Qussai Yaseen. 2021. "Spam Email Detection Using Deep Learning Techniques." Procedia Computer Science. https://doi.org/10.1016/j.procs.2021.03.107.

Abhinav, Chode. 2022. "Spam Mail Detection Using Machine Learning." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2022.44315.

Alghamdi, Bandar, and Fahad Alharby. 2019. "An Intelligent Model for Online Recruitment Fraud Detection." Journal of Information Security. https://doi.org/10.4236/jis.2019.103009.

Anita, C. S., P. Nagarajan, G. Aditya Sairam, P. Ganesh, and G. Deepakkumar. 2021. "Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms." Revista Gestão Inovação E Tecnologias. https://doi.org/10.47059/revistageintec.v11i2.1701.

Baraneetharan, E. 2022. "Detection of Fake Job Advertisements Using Machine Learning Algorithms." September 2022. https://doi.org/10.36548/jaicn.2022.3.006.

Bitenc, Marko, Tanja Cufer, Izidor Kern, Martina Miklavcic, Sabrina Petrovic, Vida Groznik, and Aleksander Sadikov. 2022. "Real-Life Long-Term Outcomes of Upfront Surgery in Patients with Resectable Stage I-IIIA Non-Small Cell Lung Cancer." Radiology and Oncology 56 (3): 346–54.

Buntain, Cody, and Jennifer Golbeck. 2017. "Automatically Identifying Fake News in Popular Twitter Threads." 2017 IEEE International Conference on Smart Cloud (SmartCloud). https://doi.org/10.1109/smartcloud.2017.40.

Chiraratanasopha, Boonthida, and Thodsaporn Chay-intr. 2022. "Detecting Fraud Job Recruitment Using Features Reflecting from Real-World Knowledge of Fraud." Current Applied Science and Technology. https://doi.org/10.55003/cast.2022.06.22.008.

G. Ramkumar, R. Thandaiah Prabu, Ngangbam Phalguni Singh, U. Maheswaran, Experimental analysis of brain tumor detection system using Machine learning approach, Materials Today: Proceedings, 2021, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.01.246.

George, Darren, and Paul Mallery. 2021. "IBM SPSS Statistics Processes for Mac." IBM SPSS Statistics 27 Step by Step. https://doi.org/10.4324/9781003205333-3.

Habiba, Sultana Umme, Md Khairul Islam, and Farzana Tasnim. 2021. "A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques." 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). https://doi.org/10.1109/icrest51555.2021.9331230.

Jakobsson, Markus, and Steven Myers. 2006. Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft. John Wiley & Sons.

Johri, Prashant, Jitendra Kumar Verma, and Sudip Paul. 2020. Applications of Machine Learning. Springer Nature.

Kontsewaya, Yuliya, Evgeniy Antonov, and Alexey Artamonov. 2021. "Evaluating the Effectiveness of Machine Learning Methods for Spam Detection." Procedia Computer Science. https://doi.org/10.1016/j.procs.2021.06.056.

Kumar M, M., Sivakumar, V. L., Devi V, S., Nagabhooshanam, N., & Thanappan, S. (2022). Investigation on Durability Behavior of Fiber Reinforced Concrete with Steel Slag/Bacteria beneath Diverse Exposure Conditions. Advances in Materials Science and Engineering, 2022.

O'Donnell, Nicola, Rose-Marie Satherley, Emily Davey, and Gemma Bryan. 2023. "Fraudulent Participants in Qualitative Child Health Research: Identifying and Reducing Bot Activity." Archives of Disease in Childhood, January. https://doi.org/10.1136/archdischild-2022-325049.

Ravindran, Pradeep Prabakar, Aditi Mishra, Prabakaran Kesavan, and S. Mohanavalli. 2010. "Randomized Tag Recommendation in Social Networks and Classification of Spam Posts." 2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA). https://doi.org/10.1109/Vbasna.2010.5730294.

S. G and R. G, "Automated Breast Cancer Classification based on Modified Deep learning Convolutional Neural Network following Dual Segmentation," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 1562-1569, doi: 10.1109/ICESC54411.2022.9885299.

Soman, Saini Jacob, and S. Murugappan. 2014. "A STUDY OF SPAM DETECTION ALGORITHM ON SOCIAL MEDIA NETWORKS." Journal of Computer Science. https://doi.org/10.3844/jcssp.2014.2135.2140.

Vaughan, Lee. 2023. Python Tools for Scientists: An Introduction to Using Anaconda, JupyterLab, and Python's Scientific Libraries. No Starch Press.