International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 12 | December - 2022 Impact Factor: 7.185 ISSN: 2582-3930

A Review on Automated Counter Measures to Filter Radicalized Content from Social Media Applications

Diksha Tatawat[1], Prof. Uday chourasia[2], Prof. Manish Kumar Ahirwar[3]

[1]Student, Department of Computer Science, University Institute of Technology, RGPV, Bhopal.

[2]Associate Prof., Department of Computer Science, University Institute of Technology, RGPV, Bhopal.

[3]Associate Prof., Department of Computer Science, University Institute of Technology, RGPV, Bhopal.

ABSTRACT: With advancements in technology and cyber warfare, terrorist and radical content is being used as a tools to spread violence and social unrest. Due to the complexity and the largeness of the data size, it is humanly infeasible to analyse the data manually or use statistical techniques. Thus machine learning based approaches are indispensable for the detection and classification of the radical and possible terrorist activity based content. In this approach, the use of social media messages in the form of tweets and messages have been considered. This paper presents a comprehensive review on filtering out potentially radical social media content based on machine learning approaches.

Keywords: Social Media, Radicalized content, machine learning, accuracy

INRRODUCTION

With the advent of social media, social engineering has also seen a tremendous upscale. Social engineering may be defined as carrying out malicious activities based on human interaction on social media platforms [1]. Social engineering has been shown to substantially jeopardize and/or manipulate human opinions and perspectives [2]. As the amount of data floating on social media platforms is staggeringly large, manual analysis of such copious amounts of data is highly infeasible. Hence, machine learning based approaches are used for the analysis of such data.Figure 1. depicts the typical message magnitude on different social media platforms [3].

Application	Per Second	Per Day	Per Month
WhatsApp	636 (thousand)	55 (billion)	1.6 (trillion)
Telegram	175 (thousand)	15 (billion)	450 (trillion)
Facebook	2.5 (thousand)	216 (billion)	6.5 (trillion)
Twitter	5.8 (thousand)	500 (billion)	15 (trillion)
Instagram	1 (thousand)	95 (billion)	2.8 (trillion)

Fig.1 Message content of Different Social Media Applications

The rampant use of social media has led to the increase in spread of all types of misinformation. The freedom of using the internet has also opened the ways towards using it in many wrong ways. The spread of radical content is one of

the most harmful ways of creation of political turbulence [4]. The proliferation of the social network on the web has given rise to a number of possibilities to share and spread any type of content. The benefit of anonymity of dark web and freedom also gives the extremist groups the advantage of doing radicalized activities on the internet through social media engagement [5]. This helps these extremists and terrorist organizations to lure and influence common people into joining their organizations and promote radically motivated acts. This also leads to politically controversial activities and increases the participation and growth of such unlawful and terror acts [6].

Radicalized information that is spread through various online mediums can endanger the security of a nation. Henceforth, it is of enormous importance to check and prevent this kind of radicalization through various social media. For this purpose, correct and accurate identification of radical content is important and proper classification is necessary. As the social media contains a huge amount of information with a huge user base, hence it is important to use a artificial intelligence and machine learning (AI & ML) based methods for identifying and classifying online radical content [7].

LITERATURE REVIEW

This section presents a brief summary of the important research findings in the domain of identifying radical content as potential social engineering outcomes.

Kapitoanov et al. in [8] proposed the Naïve Bayes classifier to filter radical content. Sadiq. et al. in [9] proposed a combination of Convolutional Neural Network (CNN) and bidirectional LSTM (CNN-bi-LSTM) for automated filtration of aggressive content on Twitter for Arabic languages. Asif et al. in [10] presented the linear support vector classifier (SVC) model for filtering extremist content from Facebook datasets using lexicon features. Fraiwan et al. in [11] proposed the SVM-OAA (one against all)and SVM-AAA (all against all) methods for classifying radical content using semantic lexicons and emotion features. The major focus of this paper was on Arabic datasets flooded by ISIS and sister organizations.

Owoeye et al. [12] proposed the use of support vector machine (SVM) as the technique to identify and classify potential radical content on social media platforms.

Rosewelt et al. [13] proposed a convolutional neural network based approach for semantic analysis of social medial tweets. The semantic analysis was used for filtering out radical content and the evaluation of the proposed work was done in terms of classification accuracy.

Lansley et al. [14] proposed fuzzy logic and decision trees based machine learning approaches for identifying radical content. Natural language processing (NLP) has been used on the semi-synthetic dataset.

Hitest et al. [15] proposed a word to vector based random forests algorithm for sentiment analysis for election result prediction and opinion polls. It was shown that the proposed work improved upon the performance of BOW and TF-IDF algorithms.

Padmaja et al. [16] proposed the use of adaptive neuro fuzzy inference systems (ANFIS)and genetic algorithm for sentiment classification of twitter data. The data sets used were twitter-sanders-apple 2 dataset. It was shown that the proposed approach beats baseline techniques in terms of positive and negative emotion classification.

Katarya et al. [17] proposed the use of Genetic Algorithm (GA) for the classification of public sentiments based on positive negative and neutral classes. Kamath et al. [18] presented an analysis of various machine learning and deep learning based approaches for sentiment classification such as decision trees, logistic regression, support vector machine, Naïve Bayes and Convolutional Neural Networks. P.Gupta et al. [22] presented the approach proposed a SVM and RF algorithm for sentiment analysis .on twitter data.

S.Agarwal et al. [23] presented The approach proposed Support vector machine technique for analysis on twitter data. M. Ashcroft et al. [24] proposed the use of support vector machine (SVM) and Adaboost as the technique. L. Kaati et al. [25] proposed the Adaboost (adaptive boosting)

algorithm. On Arabic and English tweets. M. Nouh et al. [26] proposed a RF algorithm on textual data. W. Sharif et al. [27] proposed the approach used Support vector machine based technique.. S. Mussiraliyeva et al. [28] proposed the use of support vector Machine (SVM), RF, MNB and LR algorithms. E. Ferrara et al. [29] Random forest algorithm used for sentiment analysis.

Table.1 Summary of Literature Review

S.No.	Authors	Algorithm	Feature Selection	Performance Metric
1.	Kapitoanov et al.	The approach uses the machine learning based approach using the Naïve Bayes Classifier for identifying radicalized content for twitter data.	Sentiment Lexicons	Accuracy of 89% achieved.
2.	W. Sharif et al.	The approach used Support vector machine based technique.	N-gram TF-IDF	Accuracy of 0.84 achieved.
3.	S. Mussiraliyeva et al.	The approach presented the support vector Machine (SVM), RF, MNB and LR algorithms.	TF-IDF,Word2 Vec.	Accuracy (SVM = 61%, RF = 83%, MNB =81%, LR=70%)
4.	L. Kaati et al.	The approach proposed the Adaboost(adaptive boosting) algorithm. On Arabic and English tweets.	Data dependent and data independent features	Accuracy of Arabic Tweet = 0.86, Accuracy of English Tweet = 0.99.
5.	M. Ashcroft et al.	The approach proposed the use of support vector machine (SVM) and Adaboost as the technique	Stylometric, time based, $ sentiment based.	SVM (Accuracy= 0.97%) Adaboost(Accuracy=100%)
6.	M. Nouh et al.	This approach proposed a RF algorithm.	Textual, Psychlogical $ behavioural features	FI Score = 1.0
7.	S.Agarwal et al.	The approach proposed Support vector machine technique for analysis on twitter data.	Religious war-related offensive words negative emotions and internet slang.	FI Score = 0.83 Accuracy = 0.97
8.	P.Gupta et al.	The approach proposed a	Stylometric $ Time	Accuracy of SVM = 98.03,

	(15)	SVM and RF algorithm for sentiment analysis.on twitter data.	Based features.	Accuracy of RF =98.43
9.	Owoeye et al. (12)	The approach proposed the use of support vector machine (SVM) as the technique to identify and classify potential radical content	Pro-extremist, neutral and anti-extremist content on web pages.	Accuracy = 94%
10.	E.Ferrara et al.	Random forest algorithm used for sentiment analysis	Time-based Profile, and network.	ROC_AUC = 0.93

RESEARCH FRAMEWORK

The proposed work aims at filtering out radical content and classify testing samples as potentially radical or un-radical. This however is challenging owing to the fact that there exists no clear demarcation among positive, negative and neutral data which is mined [19].

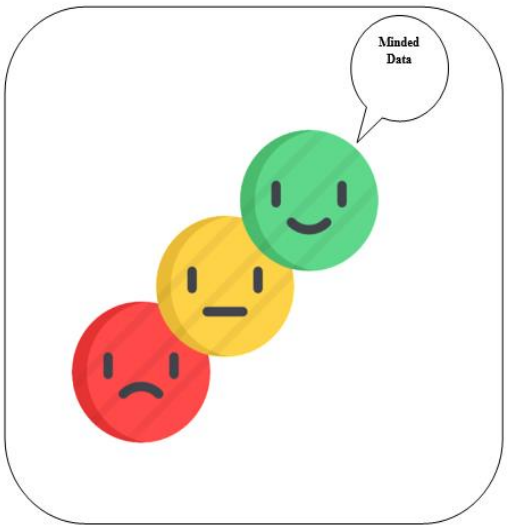

Fig.2 Sentiment Analysis

Figure 2 illustrates the concept of sentiment analysis from mined data. Moreover, merely picking up some words from a text can lead to extremely wrong calculations. For instance, Kill the President and Kill the Snake have completely different meanings. This work presents a probabilistic machine learning based approach using the artificial neural network for detection of potential radical content.

Artificial Neural networks possess a great ability to extract meaningful information from complicated data; which can applied for extraction of various features from the radical data. Some prominent traits of the ANN comprises of the following:-

1. Adaptive method of learning: An ability to sort out some way to deal with data considering the data given for planning or starting experience.
2. Way of Self-Organization: An ANN can make its own depiction of the information it gets during learning time.
3. Operation in Real Time: ANN estimations may be finished in real time that yields high accuracy and needs lesser amount of time for complex computations.

The mathematical model of the artificial neural network is depicted in figure 3.

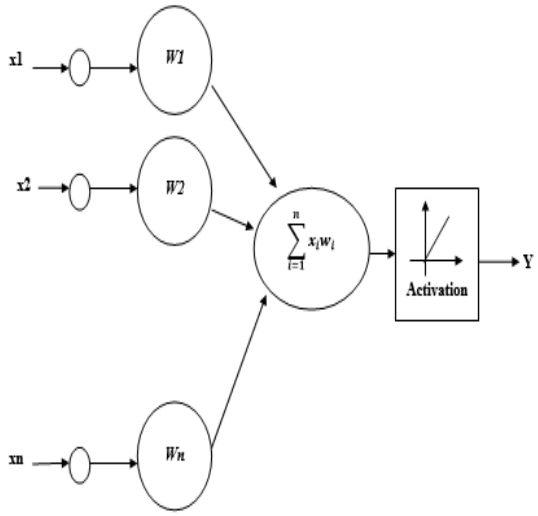

Fig.3 Mathematical model of ANN

The output of the neural network can be related to the inputs as:

$$output = f[\sum_{i=1}^{n} x_i w_i + \theta] \quad (1)$$

Here,

Output vector correcsponds to the output vector of the network.

The inputs and weights are represented by x and w respectively.

The additional term of bias termed as θ is added.

The proposed algorithm can be expressed as:

Step.1: Obtain benchmark labelled dataset.

Step.2: Divide data set into traiing and tetsing samples.

Step.3: Fix maximum number of iterations.

Step.4: Design neural network and initalize weights randomly.

Step.5: Start training and update weights.

Step.6: Check condition:

Maximum iterations over?

No: Present next training vector and increment iteration number.

Yes: Check cost function stability

The cost function is the mean squared function defined as:

$$mse = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2 \quad (2)$$

Step.7: Compute the error gradient Jacobian as:

$$J = \begin{bmatrix} \frac{\partial^2 e_1}{\partial w_1^2} & \cdots & \frac{\partial^2 e_1}{\partial w_m^2} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 e_n}{\partial w_1^2} & \cdots & \frac{\partial^2 e_n}{\partial w_m^2} \end{bmatrix} \quad (3)$$

Here,

The error e is computed as:

$$e = (y_i - y'_i) \quad (4)$$

Here,

y_i is regression target

y'_i is regression output

Step.8: Updated weights as [20]:

$$w_{k+1} = w_k - \alpha \frac{\partial e}{\partial w} \quad (5)$$

Here,

w_{k+1}&w_k denote next and present iterations respectively.

α is termed as the learning rate of the algorithm.

Step.8 Compute classification accuracy and evaluate confusion matrix based on true positive, true negative, false positive and false negative values.

As there is no clear demarcation among positive and negative word tokens for sentiment analysis, hence a probabilistic classification may yield higher accuracy. The multiclass bifurcation or division of the datasamples based on the probabilitsic approach can be given by [21]:

$$P_{Class} = Max\ [P\left(\frac{z}{c_1, c_2 \ldots\ldots c_n}\right)] \quad (6)$$

Here,

P represents probability.

P_{Class} denotes the probability to belong to class 'C'.

$P\left(\frac{z}{c_1, c_2 \ldots\ldots c_n}\right)$ denotes the conditional probability for a data sample to belong to a particular class among multiple defined classes in the set U.

A probabilistic clasifier can be effective as there exists no clear demarcation among the radical and non-radical semnatic tokens and data sets may have overlapping boundaries.

CONCLUSION

It can be concluded that with the emergence of social media, people are able to share everything with perfect freedom whenever they want and how ever they would like to. Intentional use of radicalizing the sentiments of people using online malicious strategies has been adopted by many extremist groups. Terrorist groups use the online medium to their advantage to spread radical content among individuals. Social media has become the go to platform for carrying out these unethical extremist activities by targeting large amount of online audience. It is mandatory of accurately identify potentially radical content from social media feeds. This paper presents a probabilistic neural classifier with the objective of successful identification of potentially radical content with high classification accuracy. A comprehensive review on the contemporary machine learning and deep learning based approaches has been cited so as to render insight into future directions of work.

References

1. K Chetioui, B Bah, AO Alami, A Bahnasse, "Overview of Social Engineering Attacks on Social Networks", Procedia Computer Science, Elsevier 2022, vol. 198, pp.656-661.
2. A Pimentel, KF Steinmetz, "Enacting social engineering: the emotional experience of information security deception" Crime, Law and Social Change, Springer, 2022, vol. 77, pp.341–361.
3. A. I. Kapitanov, I. I. Kapitanova, V. M. Troyanovskiy, V. F. Shangin and N. O. Krylikov, "Approach to automatic identification of terrorist and radical content in social networks messages," 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 2018, pp. 1517-1520
4. M. Nouh, J. R. C. Nurse and M. Goldsmith, "Understanding the Radical Mind: Identifying Signals to Detect Extremist Content on Twitter," 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), 2019, pp. 98-103.
5. S Davis, B Arrigo, "The Dark Web and anonymizing technologies: Legal pitfalls, ethical prospects, and policy directions from radical criminology", Crime, Law and Social Change, Springer 2021, vol. 76, pp.367–386.
6. M. Ashcroft, A. Fisher, L. Kaati, E. Omer and N. Prucha, "Detecting Jihadist Messages on Twitter," 2015 European Intelligence and Security Informatics Conference, 2015, pp. 161-164.
7. S Mussiraliyeva, M Bolatbek, B Omarov, Detection of extremist ideation on social media using machine learning techniques", Computational Collective Intelligence. ICCCI 2020. Lecture Notes in Computer Science. Vol. 12496, pp.743–752.
8. A. I. Kapitanov, I. I. Kapitanova, V. M. Troyanovskiy, V. F. Shangin and N. O. Krylikov, :A. I. Kapitanov, I. I. Kapitanova, V. M. Troyanovskiy, V. F. Shangin and N. O. Krylikov, "Approach to automatic identification of terrorist and radical content in social networks messages", IEEE Access, 2021, pp. 1517-1520.
9. S Sadiq, A Mehmood, S Ullah, M Ahmad, "Aggression detection through deep neural model on twitter", Future Generation Computer Systems, Elsevier 2021, vol.114, pp. 120-129.
10. M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, ''Sentiment analysis of extremism in social media from textual information,'' Telematics Information., Elsevier 2020, vol. 48, 101345.
11. M. Fraiwan, ''Identification of markers and artificial intelligence-basedclassification of radical Twitter data,'' Applied Computing and Information, Emerald Publication, 2020, vol. 16, no.1.
12. K. O. Owoeye and G. R. S. Weir, "Classification of Extremist Text on the Web using Sentiment Analysis Approach," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019, pp. 1570-1575.
13. A Rosewelt, A Renjit, "Semantic analysis-based relevant data retrieval model using feature selection, summarization and CNN", Soft Computing, Springer 2020, vol.24, pp.16983–17000.
14. M Lansley, F Mouton, S Kapetanakis, "SEADer++: social engineering attack detection

in online environments using machine learning", Journal of Information and Telecommunication, Taylor and Francis, 2020, vol.4, no.3, pp.346-362.

15. M. Hitesh, V. Vaibhav, Y. J. A. Kalki, S. H. Kamtam and S. Kumari, "Real-Time Sentiment Analysis of 2019 Election Tweets using Word2vec and Random Forest Model," 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), 2019, pp. 146-151.
16. K. Padmaja and N. P. Hegde, "Twitter sentiment analysis using adaptive neuro-fuzzy inference system with genetic algorithm," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 498-503.
17. R. Katarya and A. Yadav, "A comparative study of genetic algorithm in sentiment analysis," 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018, pp. 136-14.
18. CN Kamath, SS Bukhari, A Dengel, "Comparative study between traditional machine learning and deep learning approaches for text classification", DocEng '18: Proceedings of the ACM Symposium on Document Engineering, ACM 2018, Article No.14, pp.1-11.
19. M. Ebrahimi, A. H. Yazdavar and A. Sheth, "Challenges of Sentiment Analysis for Dynamic Events," in IEEE Intelligent Systems, vol. 32, no. 5, pp. 70-75.
20. R. M. Bütler, C. Häger, H. D. Pfister, G. Liga and A. Alvarado, "Model-Based Machine Learning for Joint Digital Backpropagation and PMD Compensation," in Journal of Lightwave Technology, vol. 39, no. 4, pp. 949-959.
21. X Yuan, L Xie, M Abouelenien, "A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data", Pattern Recognition, Elsevier 2018, vol. 77, pp.160-172.
22. P. Gupta, P. Varshney, and M. Bhatia, ''Identifying radical social media posts using machine learning,'' Tech. Rep., 2017, doi: 10.13140/RG.2.2.15311.53926.
23. S. Agarwal and A. Sureka, ''Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter,'' in Distributed Computing and Internet Technology. Cham, Switzerland: Springer, 2015, pp. 431–442, doi: 10.1007/978-3-319-14977-6_47.
24. M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, ''Detecting jihadist messages on Twitter,'' in Proc. Eur. Intell. Secur. Informat. Conf., Manchester, U.K., Sep. 2015, pp. 161–164.
25. L. Kaati, E. Omer, N. Prucha, and A. Shrestha, ''Detecting multipliers of jihadism on Twitter,'' in Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW), Washington, DC, USA, Nov. 2015, pp. 954–960.
26. M. Nouh, J. R. C. Nurse, and M. Goldsmith, ''Understanding the radical mind: Identifying signals to detect extremist content on Twitter,'' in Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI), Shenzhen, China, Jul. 2019, pp. 98–103.
27. W. Sharif, S. Mumtaz, Z. Shafiq, O. Riaz, T. Ali, M. Husnain, and G. S. Choi, ''An empirical approach for extreme behavior identification throughtweetsusingmachinelearning,''Appl.Sci.,vol.9,no.18,p. 3723, Sep. 2019, doi: 10.3390/app9183723.
28. S. Mussiraliyeva, M. Bolatbek, B. Omarov, and K. Bagitova, ''Detection ofextremistideationonsocialmediausingmachinelearningtechniques,'' in Computational Collective Intelligence. Cham, Switzerland: Springer, 2020, pp. 743–752, doi: 10.1007/978-3-030-63007-2_58.
29. E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, ''Predicting online extremism, content adopters, and interaction reciprocity,'' in Social Informatics (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 22–39, doi: 10.1007/978-3-319-478746_3.