International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1172-1181 www.ijaem.net ISSN: 2395-5252

K-Nearest Neighbor Algorithm to Detect Fake Identities Bots vs. Humans

[1]Versha Dadhore, [2]Prof. Nidhi Dubey
[1 2] *Department of Computer Science & Engineering*
Bhopal Institute of Technology & Science, Bhopal

Submitted: 10-12-2021 Revised: 20-12-2021 Accepted: 24-12-2021

ABSTRACT--Everyone's social life has gotten increasingly entwined with online social networks in the current age. These websites have revolutionized the way we interact with one another. Making new acquaintances and staying in touch with them has gotten more convenient. However, as a result of their quick expansion, a slew of new issues have arisen, including fraudulent pro les and online impersonation. There is currently no viable solution to these issues. We developed a methodology in this thesis that allows for the accurate and efficient identification of bogus pro less. To categorize the pro less into false or real classes, this framework employs classification algorithms like as KNN, SVM, Nave Bayes, and Decision trees. Because this is an automatic detection approach, it can be readily used by online social networks (like Twitter), which have millions of users and can't be personally checked.

False identities are a critical component of sophisticated threats that are also involved in other hostile actions. The current study examines the state-of-the-art research on identifying phony profiles on social media using a literature review. There are two types of techniques to identifying fraudulent social media accounts: those that analyze individual accounts and those that capture coordinated activity over a large number of accounts. The paper discusses the importance of phony identities in advanced persistent threats, as well as the methods for identifying fraudulent social media accounts that have been described. As a result, an appropriate KNN algorithm will be used to conduct the study for identifying bogus accounts utilizing Machine Leaning.

Keywords: Social Network Analysis, Social Media, Fake Profiles, False Identities.

I. INTRODUCTION

A human being's identity is a distinct item tied to him or her. The name of a person is a typical example. A passport, for example, comprises the person's name, birth date, and place of birth, nationality, digitally acquired fingerprints, and a digitally saved and preserved image. A third example is a Public Key Infrastructure (PKI) that uses private and public keys. In general, each identifying item should only relate to one individual. A comparable person could have several personas at the moment, such as an international ID and a few keys mentioned above, or a government disability number. Experts from many countries attest to the character's authenticity.

A cutting edge visa is an ordinary case of this. Experts ensure that the image, fingerprints, name, birthdate and so forth have a place with a similar individual, for example ensure the item connection. At an online networking webpage a client is normally recognized by a profile. It normally contains an image and name, perhaps a location and birth date. The destinations don't, be that as it may, thoroughly watch that the individual with the character implied in the profile truly made and controls the profile. On the off chance that this isn't the situation, someone is utilizing another person's character. This is called false personality.One can likewise make profiles that can utilize unreservedly designed names and other data that can't be joined to any genuine individual in any nation. For this situation the character is known as a **Faked Personality**.

Such a profile can at present contain an image of a genuine individual, picked for example haphazardly from the Internet. False characters assume a significant job in cutting edge endured dangers, for example facilitated, enduring, complex endeavors at trading off focuses in administrative, non-legislative, and business associations. False characters are likewise regularly associated with different malignant exercises, such as spamming, misleadingly expanding the quantity of clients in an application to advance it, and so forth. A normal situation for utilizing false personalities is utilizing

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1172-1181 www.ijaem.net ISSN: 2395-5252

web based life stages to imitate somebody or make a phony character to set up trust with the objective, which is then abused.

This thesis suggested a Machine Learning Algorithm for detecting impostor identities (with a focus on Twitter). Deep learning is an AI application that allows computers to learn and develop without being explicitly designed. Machine learning is concerned with the development of computer algorithms that can independently access data and learn from it. Learning begins with observations or data, such as examples, direct experience, or teaching, so that we may seek for patterns in data and make better judgments in the future will depend on our examples.

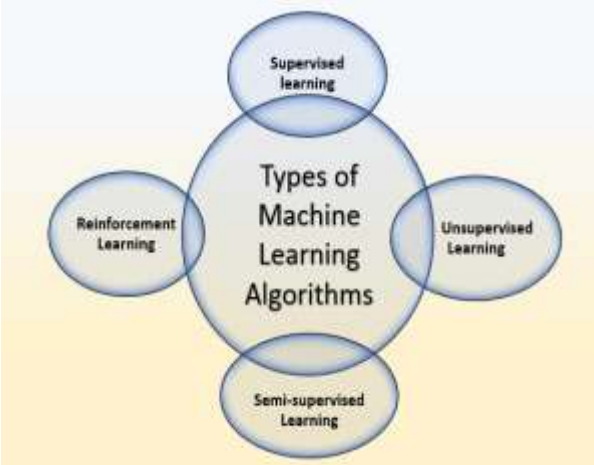

Figure 1: Machine Learning.

• Obtaining further information for a lance phishing assault, mounting a lance phishing attack, or lawfully associating to obtain intriguing data In the next section, we discuss records that were initially true but were later traded off as fraudulent records.

We also consider records to be false if they contain personal information that has nothing to do with the person who created them. It is referred to be a falsified record if the record has imagined unique peculiarities. Items that are used as IDs must be certified by specialists in the issue country, as well as viewed inside the country and beyond its borders with widespread agreement.

1.1 Social Impact

In this day and age, everyone's social life is becoming increasingly entangled with online social networks. These websites have transformed the way we engage with one another. Adding new friends and staying in touch with them has become lot easier.

Online social networks have an impact on science, education, grassroots organizing, employment, business, and other disciplines. Researchers have investigated these online social networks in order to better understand how they effect people. Teachers can instantly contact students through this, creating a welcoming environment for them to learn in. Teachers are growing more familiar with these sites, creating online classroom pages, assigning homework, holding debates, and so on, which has a big impact on education. Employers can use these social networking sites to hire people who are skilled and excited about their jobs, and their background checks can be conducted swiftly. The majority of OSNs are free, however some demand a membership fee for commercial purposes, and the remainder rely on advertising to sustain themselves. The government may use this to quickly get citizen comments.

Sixdegrees.com, The Sphere, Nex-opia (which is popular in Canada), Bebo, Hi5, Facebook, MySpace, Twitter, LinkedIn, Google+, Orkut, Tuenti (which is popular in Spain), Nasza-Klasa (which would be popular in Poland), Cyworld (which is popular in Asia), and others are some of the popular social networking sites.

II. LITERATURE SURVEY

A variety of ways to detecting fake accounts concentrate on analyzing individual social network profiles with the goal of determining the traits (or a combination of them) that assist distinguish between authentic and false accounts. Specifically, numerous information from profiles and posts are retrieved, and machine learning methods are employed to create a classifier capable of recognizing bogus accounts (Table 1).

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1172-1181 www.ijaem.net ISSN: 2395-5252

In the research Nazir et al. (2010), for example, apparition profiles are recognized and depicted in online social gaming apps. The study investigates a Facebook application, the web-based amusement "Warriors club," which is acknowledged for providing motivating forces and a gaming favored perspective to clients who welcome their looks into the game. According to the authors, providing such motivators encourages gamers to create false accounts. The client's motivation esteem would grow as a result of bringing those false profiles into play. The Authors begin by removing 13 highlights for each game client, and then utilize support vector machines to aggregate them (SVMs). These tactics, according to the article, do not offer any irrefutable discriminants between real and counterfeit clients [6].

Fake profiles in LinkedIn are identified by Adikari and Dutta (2014). Using restricted profile information as information, the article shows that fraudulent profiles can be identified with 84% accuracy and 2.44 percent false negative. Neural networks, SVMs, and top-of-the-line segment analysis are all linked. Features such as the number of dialects spoken, education, aptitudes, recommendations, hobbies, and grants are used, among other things. As a ground truth, attributes of known-to-be-fake profiles uploaded on unique websites are used.

Chu et al. (2010) advocate for distinguishing human, bot, and cyborg Twitter accounts (i.e., bots and people working in show). An Orthogonal Sparse Bigram (OSB) content classifier that uses sets of words as highlights is used as part of the discovery problem plan to identify spamming/fake accounts accounts. The system was able to distinguish between bots and human-worked accounts using additional recognizing components such as the normalcy of tweets and other record attributes such as the recurrence and types of URLs, and the use of APIs.

Lee et al. focused their work on identifying bogus accounts on Twitter, exactly as they did on MySpace (2010). In comparison to Chu et alstudy, .'s the list of highlights in this study was expanded to include the number and kind of relationships. The Decorate meta classifier was determined to offer the greatest order exactness out of the several classifiers available in the Weka AI package.

Regardless, or rather than breaking down individual profiles, a new wave of approaches rely on chart-based highlights to distinguish between fake and genuine data. Stringhini et al. (2010), for example, discuss Facebook and Twitter efforts for detecting false IDs. For a year, the authors created 900 honeypot profiles in informal networks and collected approaching messages and companion needs on a continuous basis. Client information from those who carried out the solicitations was gathered and analyzed, and around 16K bogus accounts were discovered. The authors went on to investigate how artificial intelligence may be used to locate false accounts. The Authors used message comparability, the proximity of instances behind the search for companions to include, and the proportion of companion solicitations in addition to the features used in the previous studies, and then used Random Forest as a classifier.

Table 1: Profile-based methods for detecting fake social media accounts.

Reference	Ground truth	Detection method
Adikari 2015	Fake LinkedIn profiles have been discovered on dedicated websites.	Characteristics such as the number of languages spoken, education, abilities, recommendations, interests, and accolades are used to train neural networks, SVMs, and principal component analysis.
Chu et al. 2010	3000x2 Twitter accounts were manually classified as humans, bots, or cyborgs.	Humans, bots, and cyborgs were manually categorised as humans, bots, and cyborgs on 3000x2 Twitter accounts.
Lee et al. 2010	Honeypots create fictitious accounts: On MySpace, there are 1500 people, and on Twitter, there are 500 people.	We put over 60 Weka classifiers to the test. Demographics, content quality, and frequency of content creation, as well as the quantity and kind of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1172-1181 www.ijaem.net ISSN: 2395-5252

		connections, are all factors to consider. The Decorate meta-classifier provided the best results.
Stringhini et al. 2010	Honeypots create fictitious accounts: There are 173 bogus Facebook profiles and 361 fraudulent Twitter accounts.	The random forest was constructed using the ratio of accepted friend requests, URL ratio, message similarity, regularity in friend selection, messages sent, and number of friends.
Yang et al. 2011a	Fake Twitter accounts are classified as those with harmful URLs: 2060 fictitious accounts	Graph-based features (local clustering coefficient, between centrality, and bi-directional links ratio), neighbor-based features (average neighbors' followers), automation-based features (API ratio, API URL ratio, and API Tweet similarity), and timing-based features were all used to build different classifiers.

Some extant recent algorithms are reviewed in the comparison table 1 above, along with their benefits, drawbacks, limitations, and potential future extensions.

III. PROPOSED METHODOLOGY

With the following data related with the extension used to find a defined document false, an ANN-based technique is offered to work with problem-solving challenges. The link between actual and objectionable ratings is examined in our approach to fake-Id issues. Names, pronouns, advertisement, adverb, verb, determinator, coordination of conjugations, prepositions, and predicates were used to create 9 function values for each validation based on the POS tags. These POS tag features serve as a benchmark for comparing traditional model creation with various algorithmic approaches.

3.1 Steps to a Successful Execution

The k-closest neighbor (K-NN)classifier is an excellent illustration classifier, which implies that the prepared archives are utilized for correlation rather than a clear categorization description, as is the case with other classifiers' classification profiles. In that position, there isn't any actual training. When a new report has to be sorted, the k closest records (neighbors) are found, and if a sufficient number of them have been allocated to a certain classification, the new archive is allocated to that classification as well. Furthermore, the nearest neighbors can be found using standard questioning methods. To assess if a message is legitimate or not, we examine the class of the messages that are closest to it. The method of joining the vectors has been used for many years. The k nearest neighbor computation's probability is as follows:

Preparation **(stage 1)**

Preparation messages should be saved.

Sifting, **stage 2**

Select the k closest Neighbors of a message x from the readiness set of messages. If these neighbors have additional bogus IDs, classify the message as phony. Authentic mail is the most common classification. It's worth noting that using an ordering technique to shorten the evaluation period results in a model with a multidimensional nature O(m), where m is the previous measure. This approach is also proposed as a memory-based classifier since the majority of the planning points of reference are stored in memory. Another issue with the calculated results is that there appears to be no parameter that could be adjusted to reduce the number of false positives. By modifying the layout guideline to the following l/k rule, the problem is effectively resolved:

If l or more of x's k nearest neighbors' messages are phony, classify x as fake and send it as real mail.

When it comes to arranging assignments, the k nearest neighbor rule has gained a lot of popularity. It's also one of the few consistent characterization rules in the world.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1172-1181 www.ijaem.net ISSN: 2395-5252

3.2 Algorithm Pseudo Code:

Algorithm: K-NN Algorithm for Outlier Detection
Input: Applications, Threshold value, Fake-Id detection System framework installation
Output: Fake-Id app listing, computation parameter, utilization monitoring.

Steps:

```
Begin [
Loading Configuration framework;
Loading Fake-Id Info();
int n=numofapp();
foreach(1;n)
{
finding app usage();
finding http request monitor();
data Fake-Id analysis();
permission usage analysis();
Detection usage per unit of time();
foreach app()
{
finding its statistic usage();
stack analysis();
}
computing optimal value()
{
finding app usage();
optimal data use();
Fake-Id analysis();
returnEv(Energy value);
}
if(Ev>th Value)
{
add Vector Listing();
}
return Intrusion app listing();
}
return computation parameters();
]
End;
```

The Figure below represents the complete flow of the proposed scenario which represents our work and computes parameters efficiently.

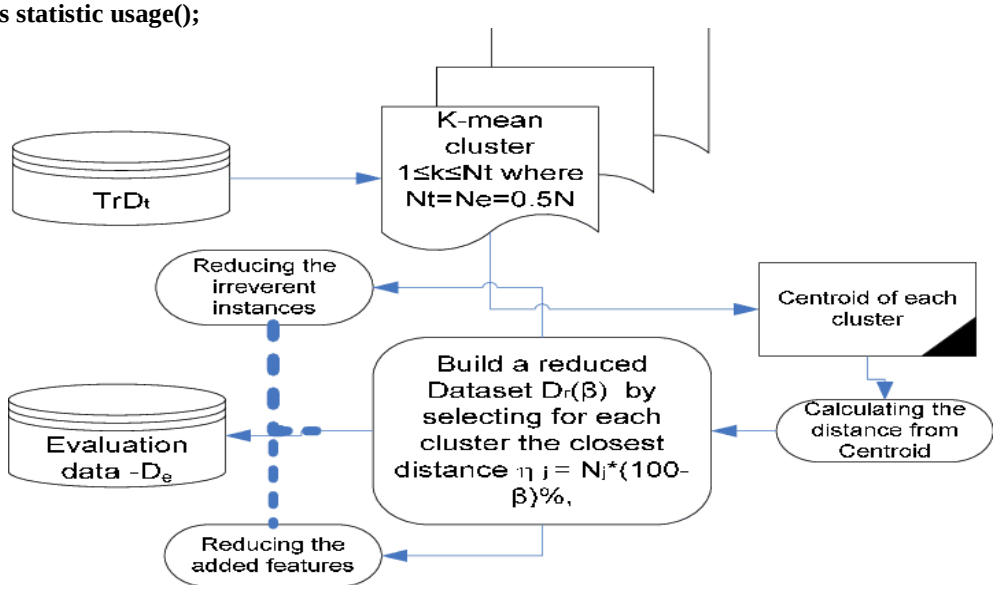

Figure 2: Flow chart of Fake-Id Detection using K-NN

In this paper, we define the problems of the existing default algorithm in the scenario and define the intent of our working definition. For example, the high-efficiency work of our President has additionally led to a work research record for the predictive model, querying input for system processing and outperformance parameterization.

IV. SIMULATION ENVIRONMENT

We're working on a two-factor authentication application right now. We used JDK 1.8, which is a Java development kit, to create this program. JAVA is a programming language that is used to create computer programs. It aids the programmer in writing computer commands in English. Because it is reliable and easy to write, this form of language is known as high-level language. The way instructions are written in JAVA is determined by a set of rules. Syntax is the term for these rules. The high-level instructions of a program are converted into numeric codes that computers can comprehend and execute once it has been written. Enterprise software and content on the web. The JAVA development kit (SDK) is a software development kit (SDK) for writing JAVA programs. Oracle's Java soft division develops the JDK.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1172-1181 www.ijaem.net ISSN: 2395-5252

The technique is built on a java platform with 8 GB of RAM and 1 TB of hard drive, utilizing the pubnet dataset of documents, which contains various documents relevant to Fake-Id content. Swing & Chart Api-based implementation is used to execute the analysis. The result is computed in milliseconds, and the similarity measure is computed in percents. The following is the outcome of utilizing KNN-ROBKP, which is a KNN-based robin karp Fake-Id detecting technique. When two parameters are compared, the technique determines its efficacy.

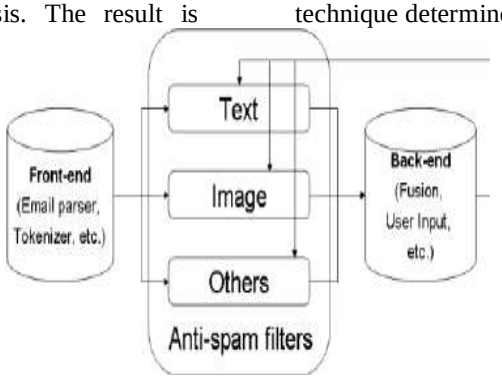

Figure 3: Block diagram of Fake-Id filtering

4.1 Dataset

Recruitment Post – In order to perform the experimental setup and result in the analysis part requirement is to access a large dataset, We created our own dataset which contains posts with dates for the analysis. We created our dataset such that it comprises of a mixture of posts such as general posts (text from a normal user) as well as posts with vulnerable content (text from a violent group). Then we tried to categorize the dataset into Extreme Violation Detected---YES or NO.

4.2Observation

The results were performed on the same dataset as discussed above and both the methods were applied over similar dataset so as to get comparative results. We created our own dataset which contains posts with dates for the analysis. We created our dataset such that it comprises of a mixture of posts such as general posts (text from a normal user) as well as posts with vulnerable content (text from a violent group). Then we tried to categorize the dataset into Extreme Violation Detected --- YES or NO…

Table 2: Comparison analysis obtained through Traditional Fake-Id filtering method vs. KNN-ROBKP proposed scenario.

Algorithm	Computation time in msec.	Similarity measure
Traditional method	86.72	50.00
KNN-ROBKP	68.72	64.32

The above table 2, discuss about the computation observed during the result computation.

Comparison analysis graphical representation:-

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1172-1181 www.ijaem.net ISSN: 2395-5252

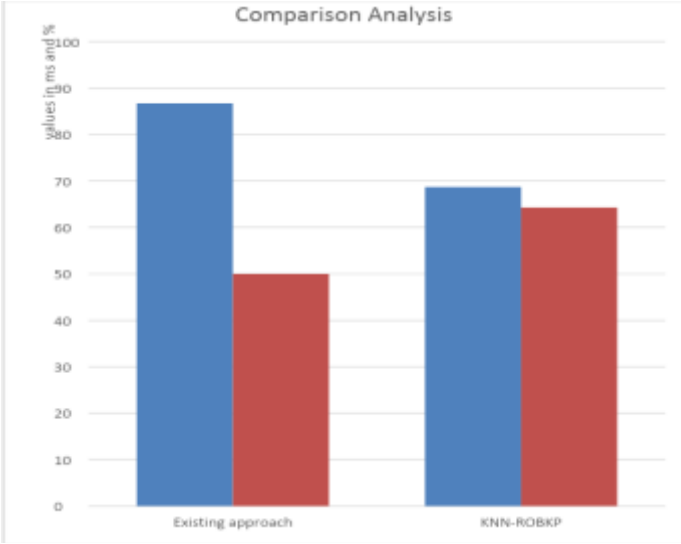

Figure 4: Comparison analysis graphical representation.

The comparison and efficiency of the suggested method utilizing the supplied parameter is depicted graphically in figure 4 above. As a result, the suggested technique is quite beneficial for doing Fake-Id analysis on data using the KNN-ROBKP algorithm.

V. CONCLUSION & FUTURE WORK

On the one hand, technology provides us with several benefits; on the other hand, a problem occurs with the usage of email, which is that trash mail may be sent to anyone's email address by a third party, and that mail will be classified as Fake-Id messages or garbage mail. Five machine learning algorithms for Fake-Id screening are proposed in this study. As a result, the study provides a general review of spa filtering methods. Fake-Id communications are unsolicited and cannot be specified in any way. It causes a slew of issues in the economic and ethical spheres, leading to legislative measures to define and outlaw Fake-Id. KNN-classifier based filtering is a method that is highly recommended and used to detect Fake-Id. Currently, a variety of techniques are used to categorize various types of emails according to certain criteria. It usually refers to the automated processing of methoding communications, but it may also refer to the mediation of human understanding, as well as the active messages, such as those received. As a result, the proposed efficiency approach can be utilized instead of the N-gram method. Applying the proposed strategy to a real-time dataset will require more research. Using IDS approaches, the procedure can also be improved.

Fake-Id problems detecting with the N-Gram function analyzes a relatively good result and is well-suited. We see that the linguisticfeatures of the positional model provide secondary support for our classification model. The combined model provides smarter results as well as thepsychological tendencies of the spammer tablet.We also suggest that our Fake-Id without the rapper s informationdoes not analyze enough.It greatly enhances the use of our data. Some of the developing metadata, such as review reviews, amount of reviews, reviewer's IP address, rapper's age, and so on, may be quite useful in identifying bogus reviews and Fake-Id in our Fake-Id study. Unfortunately, we receive user information on the stated websites for privacy concerns, and these websites are the only ones that can analyze user data internally. We might also use ratings of information kept on official sites for stored items, such as electron guns vs counter-addresses, or tech-crunch and hotel releases against review critics, to find the genuine quotient of the content. Nonetheless, we might utilize the presented work as a starting point for further study in this field.

REFERENCES

[1]. S. Gu rajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, “Profile characteristics of fake twitter accounts,” Big Data & Society, vol. 3, no. 2, p. 2053951716674236, 2016.

[2]. C. Xiao, D. M. Freeman, and T. Hwa, “Detecting clusters of fake accounts in o nline social networks,” in Proceedings of the 8th ACM Workshop on Artificial

Intelligence and Security. ACM, Conference Proceedings, pp. 91 – 101.
[3]. S. Mainwaring, We first: How brands and consumers use social media to build a better world. Macmillan, 2011.
[4]. V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, "The darpa twitter bot challenge," arXiv preprint arXiv:1601.05140, 2016.
[5]. Y. Li, O. Martinez, X. Chen, Y. Li, and J. E. Hopcroft, "In a w orld that counts: Clustering and detecting fake social engagement at scale," in Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Conference Proceedings, pp. 111 – 120.
[6]. Nazir, A., Raza, S., Chuah, C.-N., Schipper, B., 2010. Ghostbusting Facebook: Detecting and Characterizing Phantom Profiles in Online Social Gaming Applications, in: Proceedings of the 3rd Wonference on Online Social Networks, WOSN'10. USENIX Association, Berkeley, CA, USA, pp. 1–1.
[7]. Adikari, S., Dutta, K., 2014. Identifying Fake Profiles in Linkedin, in: PACIS 2014 Proceedings. Presented at the Pacific Asia Conference on Information Systems.
[8]. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S., 2010. Who is Tweeting on Twitter: Human, Bot, or Cyborg?, in: Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC '10. ACM, New York, NY, USA, pp. 21–30. doi:10.1145/ 1920261.1920265.
[9]. Fayazi, A., Lee, K., Caverlee, J., Squicciarini, A., 2015. Uncovering Crowdsourced Manipulation of Online Reviews, in: Proceedings of the 38th International ACM SIGIR Conference on Paper and Development in Information Retrieval, SIGIR '15. ACM, New York, NY, USA, pp. 233–242. doi:10.1145/2766462.2767742.
[10]. Egele, M., Stringhini, G., Kruegel, C., Vigna, G., 2015. Towards Detecting Compromised Accounts on Social Networks. IEEE Trans. Dependable Secure Comput. PP, 1–1. doi:10.1109/TDSC.20
[11]. James,G.,Witten, D., Hastie,T.,& Tibshirani, R.(2013). An Introduction to Statistical Learning.
[12]. Symantec,Stateof Spam andPhishing.A Monthly Report 2010, http://symantec.com/content/en/us/enterprise/other_res ources/state_of_spam_and_phishing_report_09-2010.enus.pdf.
[13]. NehaSingh,Dendritic Cell Algorithm and Dempster Belief Theory Using Improved Intrusion Detection System, InternationalJournal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013 ISSN: 2277 128X.
[14]. EnricoBlanzieri,AntonBryl,2008.A surveyof learningbased techniques of email spam filtering, Technical Report .
[15]. JulieGreensmith,The Dendritic CellAlgorithm ,Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy October 2007.
[16]. Alsmadi,Izzat,andIkdamAlhami. Clustering and classification of email contents.Journal of King Saud University-Computer and Information Sciences27.1 (2015): 46- 57.
[17]. Shrivastava,J. N.,and Maringanti, H. B.,E-mail spam filtering using adaptive genetic algorithm, in Inter national Journal of Intelligent Systems and Applications, ,2014,Vol. 6(2),pp. 54.
[18]. Y. Tan, G. Mi, Y. Zhu and C. Deng, Artificial immune system based methods for spam filtering, in 2013 IEEE International Symposium on Circuits and Systems (ISCAS2013), Beijing, 2013, pp. 2484-2488. DOI: 10.1109/ISCAS.2013.6572383.
[19]. BahgatE.M.,RadyS.,andGadW.,(2016).AnE-mailFiltering method Using Classification Techniques,In T. Gaber,A.E. Hassanien,N.El-Bendary,& N. Dey (Eds.), The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef,Egypt Cham: Springer International Publishing.pp.321–331.DOI: https://doi.org/10.1007/978-3-319-26690-9_29.
[20]. Elssied,Nadir Omer Fadl,OTHMAN IBRAHIM,and Waheeb Abu-Ulbeh. AN IMPROVED OF SPAM E-MAIL CLASSIFICATION MECHANISM USING K-MEANS CLUSTERING. Journal of Theoretical & Applied Information Technology 60.3 (2014).
[21]. Symantec Corporation.(2016).Internet Security Threat Report (Vol. 21).
[22]. M.Erdélyi and A.A.Benczúr,Temporal analysis for web spam detection: an

overview, in Proceedings of the 1st International Temporal Web Analytics Workshop, pp. 17–24, March 2011.

[23]. Dasgupta, Anirban, Maxim Gurevich, and Kunal Punera. Enhanced email spam filtering through combining similarity graphs. Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

[24]. L.Han andA. Levenberg,Scalable online incremental learning for web spam detection, in Recent Advances in ComputerScienceandInformation Engineering: Proceedings of the 2nd World Congress on Computer Science and Information Engineering, vol. 124 of Lecture Notes in Electrical Engineering, pp. 235–241, Springer, Berlin, Germany, 2012. View at Publisher · View at Google Scholar

[25]. J.HuaandZ. Huaxiang, Analysis on the content features and their correlation of Web pages for spam detection, China Commun., vol. 12, no. 3, pp. 84–94, Mar. 2015.

[26]. N.Pérez-Díaz,D.Ruano-Ordas,F.Fdez-Riverola,andJ.R.Méndez,Wirebrush4SPAM: anovelframework for improving efficiency on spam filtering services, Software: Practice and Experience, vol. 43,no.11, pp. 1299–1318, 2013. View at Publisher · View at Google Scholar · View at Scopus

[27]. Scheltus,P.,Dorner,V.,&Lehner,F.(2013). Leave aComment! An In-Depth Analysis of User Comments on YouTube. Wirtschaftsinformatik,42.

[28]. Islam,Rafiqul,and Y.Xiang.Email classification using data reduction method. Communications and Networking in China (CHINACOM), 2010 5th International ICST Conference on. IEEE, 2010.

[29]. Hamou,R.M.,Amine, A., & Tahar, M. (2017).The Impact of the Mode of Data Representation for the ResultQuality of the Detection and Filtering of Spam. In Ontologies and Big DataConsiderations for Effective Intelligence (pp. 150-168). IGI Global.

[30]. Z.Gyöngyiand H.Garcia-Molina, Web spam taxonomy, in Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '05), pp. 39–47, Chiba, Japan, May 2005

[31]. Alsaleh,M.,Alarifi, A.,Al-Quayed,F.,&Al-Salman, A. (2016). Combating comment spam with machine learning methodes. Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015, 295–300. https://doi.org/10.1109/ICMLA.2015.192

[32]. O. Osho, V. L.Yisa,O. Y.Ogunleke, andS. I. M. Abdulhamid, Mobile spamming in Nigeria: An empirical survey, in Proc. Int. Conf. Cybersp. (CYBER-Abuja), Nov. 2015, pp. 150–159.

[33]. Sah,U.K.,&Parmar,N.(2017). An method for Malicious Spam Detection in Email with comparison of different classifiers.

[34]. D.Fetterly, M.Manasse,andM.Najork, Spam, damn spam, and statistics: using statistical analysis to locate spam web pages, in Proceedings of the 7th International Workshop on the Web and Databases (WebDB '04), pp. 1–6, ACM, Paris, France, June 2004. View at Publisher · View at Google Scholar · View at Scopus

[35]. Islam,Rafiqul,andY. Xiang. Email classification using data reduction method. Communications and Networking in China(CHINACOM).

[36]. Kaushal,R.,Saha, S.,Bajaj, P.,&Kumaraguru, P. (2016, December). KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube. In Privacy,Security and Trust (PST), 2016 14th Annual Conference on (pp. 157-164). IEEE.

[37]. S. Ghiamand A. N. Pour,A survey onweb spam detection methods: taxonomy, International Journal of Network Security & Its Applications, vol. 4, no. 5, pp. 119–134, 2012. View at Publisher · View at GoogleScholar.

[38]. Manwar,S.R., Lambhate, P.,& Patil, J. (2017). Classification Methods for Spam Detection In Online Social Network.

[39]. N.S.Kumar,D.P.Rana,R.G.Mehta, Detecting Email Spam using spam word associations, IJETAE,ISSN 2250-2459, Volume 2,Issue 4,April2012.

[40]. Massey,B.,Thomure,M., Budrevich, R., Long, S.: Learning spam: Simple techniques for freely-available software. Usenix 2003, Freenix Track(2003)

[41]. More,Mugdha and Bharattidke."Social media opinion summarization using ensemble technique." Pervasive Computing (ICPC), 2015 International Conference on IEEE2015.

[42]. Arya,S.,Mount, D.M., Netanyahu, N.S.,Silverman, R.,Wu, A.Y.: An optimal algorithm for approximate nearest neighbor

searching in fixed dimensions. Journal of the ACM 45 (1998) 891–923

[43]. WalinMa, DatTran,Dharmendra Sharma, A Novel Spam Email Detection system based on Negative Selection, 2009 Fourth International Conference on Computer Science and Convergence Information Technology.

[44]. Group,P.:The portlandspam automatic mail-filtering project (2003) Accessed 02/03/04

[45]. NitinJindal andBingLiu, Review Spam Detection,Proceedings of 16th International World Wide Web conference,WWW '07. Ban_, Alberta, Canada 2007.

[46]. Quinlan,J.R.:C4.5, Programs For Machine Learning.Morgan Kaufmann,California(1993).