P- ISSN 1991-8941 E-ISSN 2706-6703 *Journal of University of Anbar for Pure Science (JUAPS)* Open Access
2019,13 (1) :41- 49

Apriori Method of Mining Secure Data in Social Media

Zahraa Raji Mohi

Department of Advocacy and rhetoric ، Al-Imam Al-adham University College

ARTICLE INFO

Received: 13 / 5 /2019
Accepted: 18 / 8 /2019
Available online: 19/8/2022

DOI: 10.37652/juaps.2022.171873

Keywords:
Social media Data.
Facebook.
Data security.
Apriori algorithm.
RSA algorithm.

ABSTRACT

As can be seen all around us, social media sites such as Facebook, Twitter, YouTube and Flickr and other became more importance grow rapidly in the last few years. This growth in social media sites has led to increase of information generated and circulated between individuals, this information significance in the companies and institutions works and also for individuals, for this it is important to analyze and classify data by determining keywords and main sentences which lead companies to manage their works more better with present and possibly clients.

However, social media data may be contain various types of unwanted and maleficent spammer or hacker actions. So, there is a critical need in the social media network and society, industry for social media security. In this paper, we choose Apriori method for mining and classifying social media data and take a Facebook to be a case study for social media data then after classifying and mining data applying RSA algorithm which is most popular and easer to implement secure data and use it usefully in the company's work.

1- Introduction:-

Online people group have created to concentrate on both individual and expert lives. Countless gatherings have been framed concentrating on each potential ground of excitement, including feeding, sports, music, child rearing, scrapbooking, and other genuine issues. It is assessed that there are more than 900 online life locales on the web these days. Probably the most prominent stages are Facebook, Google Plus, YouTube and Twitter, where [1]:

• Every month people on Facebook go through more than 500 billion minutes of time.
• More than 200 million have Twitter accounts.
• Every day 2 billion watchers of YouTube videos.
• 77% of web users reading sites.

Most of the populace is utilizing online networking in some structure or another. Given the considerable increment in the utilization of internet based life, there is a lot of data that is being created [2]:

• 30 billion bits of substance and all the more every month is shared on Facebook.

• 24 hours of video every moment is transferred to YouTube.

• The normal number of tweets sent every day was 110 million in of December 2010.

• 133 million sites recorded on driving online journal index.

Because of the much time spent via web-based networking media and the tremendous measure of data being produced, organizations have paid attention and are endeavoring to use the intensity of web based life to enable them to succeed [3].

Since a significant part of the information from the web-based social networking destinations are content based information, the way toward planning and breaking down the information will concentrate on standards of getting ready content information for examination. Figure 1 demonstrates the proposed method of mining secure social media by taking

* Corresponding author at: Department of Chemistry, Ibn-Al-Haithem College of Education for pure science. University of Baghdad, Iraq.E-mail address: **zahraa_raji_cs@yahoo.com**

Facebook posts first collecting text data from it then processing and cleaning the data to applying Apriori mining algorithm on the data finally the RSA security algorithm will be performed on the classified data to use securely through the social media again.

1.1 Related Work:

- **Wei Dong et al (2011):** It mentioned how mobile social network extended into the cyberspace through the real world and how can we discover secure friend on the social media and protect our devices by identifying the range of potential attacks and analyzing real traces then develop novel solution for secure potential friend and finally explain feasibility and efficiency of the computation applied to develop a secure friend discovery protocol.
- **Wu he (2013):** It mentioned that mobile malware and viruses increasing day by day using social media on mobile and become a very popular attack today so the use of state of art of the security aspect of mobile social media and blog mining based on the detailed review and the author summaries some aspects to understand risks associated with social media mobile security.
- **Dan Bogdanov et al (2016):** It introduced a multi-party computation as a technique suited for privacy-preserving data mining by using Sharemind as a secure multi-party computation applied on several scenarios and large dataset then confidential processing on data and using Sharemind model for share conversation equality , bit extraction and it find the mining of large data using k-means algorithm of clustering shows that secure multi-party computation ready to handle real world by handling hundreds of millions of execution in reasonable time.
- **Kia Shu et al (2017):** It mentioned the detection of fake news on social media society by exploring review existing data using two phases 1st characterizing, 2nd detecting where the concepts and principles of fake news and using data mining perspective include feature extraction and model construction to found feature direction on fake news detection depending on traditional and social media principle of fake news.

1.2 Data Security:

There are a large number of social security services provided by each government, and each service is carried out through different steps. So, a huge amount of data is created with each service. In this paper, the simple business workflow is discussed. Because of actualizing everyday government managed savings benefits, a colossal measure of information have been amassed which increments drastically consistently. In any case, the present work is keen on Facebook information prepared by Apriori calculations that are then the dissected much effectively for standardized savings.

As of late, messages are progressively being traded over online life, which has brought about gigantic increment in the speed of correspondence. In the meantime, this has offered ascend to have new issues, for example, security dangers [9].

1.3 Data Mining:

Data mining is a procedure of consequently finding valuable data in huge information vaults that utilizes an assortment of information examination devices to find examples and connections that can be tucked away among huge measure of information [7].

Data mining career is used to indicate the array of paradigms to get information. When all is said in done, they are ordered into two classifications: (1) clear and (2) prescient. Expressive mining mission is to photo the general characteristics of the data in the database. Prescient mining mission perform deduction on the present data so as to make forecasts. Notwithstanding, content mining and database information disclosure are dealt with as often as possible as equivalent words; content mining is really part of the learning revelation process. Information mining isn't determinant to one sort of media or information; information mining ought to be somewhat feasible to a data distribution center. Be that as it may, calculations and techniques may contrast when connected to variable kinds of information [8].

1.4 Social Media Data:

The most characterizing highlight of Social Media systems is the association or the like of media, regardless of whether they are as pictures, recordings,

writings, or different structures. This media is shared, socially, among its clients. The real distinction between Social Media locales and customary media destinations is that most of the substance via web-based networking media locales is produced by its clients. This is contrary to the standard "Mass" media where almost all substance is made by a lot of distributers, who are extraordinarily by the client base. Along these lines, the measure of substance on Social Media systems is a lot bigger than conventional media arranges and develops at a rate that is just constrained by the quantity of partaking clients [4].

Online life, be that as it may, has significantly less characterized jobs of maker and buyer. As a rule, most of the client base performs the two jobs all the while. All things considered, there are numerous internet based life systems where clients can change content, re-distribute content that has either been adjusted or had remarks annexed to it, or add Meta data to content [5].

2- Analyzing Social Media data Challenges:

Breaking down web based life information picked up esteem is incredible; numerous difficulties related with online life examination which will require further investigation a portion of these difficulties are the accompanying:

1-Accessing and gathering data:- a few applications accessible are enable organizations to start gathering and investigating web based life information in addition, organizations additionally can be able to assemble programs inside that do this.

2-Analyzing content information: in commonly when a Facebook present is a replication on different posts. The data that is being replayed to might possibly be accessible agreeing on how the client is replaying or what an organization might follow.

3- If the degree isn't accessible, it will be a major test for the examination to get a handle on precisely what this data implies. When it is accessible, the correct arrangement of internet based life information associating together to most likely comprehend the more extensive setting of a discussion is turned into the test [5,1].

Facebook is an online web based life administration. It is site was propelled on February 2004 by Mark Zuckerberg with his companion in Harvard College. First Facebook authors had the site enrollment to Harvard understudies, and then it will extend it to universities in the Boston Area. It helps students at various colleges and also extended to secondary school students. From 2006 till now, any person who is more 13 years old is allowed to be as Facebook customer [Wikipedia, 2].

Over 1.44 billion on Facebook dynamic customers had spent time as of March 2015. A consequence of the expansive size of information customers transmit to the management, Facebook has gone down investigation to ensure branches [6,4].

3- Proposed Data manipulation:

The focal point of this paper is on deciding the theme matter related with a post. Words in a post are simply words. They don't convey the correct feeling that is shows up in an eye to eye discussion, in which case the individual could distinguish satisfaction, bitterness, mockery, irate, and so on. There are things that clients endeavor to attempt and changing distinctive feelings like (smiley face, dismal face, rehashing a few words commonly, "lol," TYPING IN ALL CAPS, and so forth.), however notwithstanding when an individual peruses an electronic correspondence they may get the emotion off-base.

In this paper populist pages with over 1,000,000 fans are chosen to be as basic measurers to monitor the public posting activities in the social network. One of the most popular American lifestyle clothing brand are followed by young people is Hollister. Its Facebook page has 4,608,404 fans in March 2011 and 5,100 user added photos/videos, most of which are photos. For example if we have 6 posts in the last 5 minutes it is possible that 4 of them are from spammers or hackers[3].

3.1 Collecting and processing text data Steps:

First and vital thought is that online life information will in general be casual with incorrect spellings and contractions issues will be a bigger test. Moreover, on account of Facebook, there are named images that really do have significance for this images

an additional consideration needs when a content cleaning is performed [6].

This data and images contain single and twofold citation signs, brackets, accentuation signs, and stray images (dollar signs, stars, and so forth.). In the underlying information cleaning, the signs that really have importance for Facebook (@, #) were held. In different wellsprings of online life such like Facebook where one will require a lot a bigger number of words than this to catch all the substance.

There are a few stages that can be perceived. These stages are, portrayed beneath. These stages are iterative and done before mining to filter useful data as shown in Figure1.

In data collecting stage: The information dependent on the necessities of either those coordinating the input or customers who will utilize the completed result of the investigation. Information might be numerical or explicit (i.e., content mark for numbers).

In data processing stage: That is necessary to share by experts to trustees of data, for parable, data origination action power into an association, while in data cleaning: data can be fragmented, contain copies, or contain mistakes. The necessity for information improvement can emerge from problems in however information is entered and put away. Information improvement is that the means toward obviation and redressing these blunders [1, 2].

3.2 Apriori Mining Algorithms:

Association Mining Rules find interesting associations or correlations among a large set of data items, for example: "one may discover a set of symptoms frequently occurring together with certain kinds of diseases and further study the reasons behind it" [6].

In this examination, it is expected that each post has measure up to weight. Not with standing, they cause that one should need to gauge posts contrasted. One reason is on the grounds that users have varied quantities of preferences, and a post from a user with 2,000 preferences is probably going to be seen by a greater number of users than a post from a user's with 200 preferences. So a post could have an especially weight given the effect of the individual sending the post. Another purpose behind giving a post an especially weight might be the way that it is shared by others. In the event that it is a remark that different clients are acknowledged with, it can spread quicker and sway more clients. So the quantity of preferences and offers ought to be taken in weighting posts in Table 1 we show the attributes selected for each user and the category contain, for example the user belong to all categories take higher weight .

The Apriori calculation is appeared in Figure 3.

Where the data categorized according to a predefined classes contain information about the authenticated user to take the information from them and applying the security algorithm by receive useful information .A large number of successive itemsets, expansive itemsets, or extremely low least help, regardless it ridden from the expense of producing an immense number of applicant sets and filtering the database over and over to check a huge arrangement of hopeful itemsets. Truth be told, it is important to create 2100 competitor itemsets to increase visit itemsets of size 100[7].

3.3 The RSA Algorithm:

A well-known encryption calculation is RSA and it is one of the most punctual calculations and still utilized productively. Information encryption standard treats the information obstruct in its progressively straightforward structure, a gathering of bits RSA expected that the plaintext square is a symbolizing of a whole number utilizing base-two math. Therefore, the calculation guessed that the plaintext P and the cipher text C are numbers symbolized by the information in the plaintext and the figured content squares. Moreover, the calculation presumes that these two whole numbers between the range $0 \leq P < n$, $0 \leq C < n$, where the modulus is thought to be the result of two prime (numbers without any elements other than themselves and one)[9].

The framework created by Rivest, Shamir and Adleman makes utilization of an articulation with exponentials. The numerical depiction of RSA is as per the following: take two huge primes, p and q, and process their item n = pq; n is known as the modulus. Pick a number, e, not as much as n and moderately prime to Ø (n), where Ø (n) = (p-1)(q-1). This implies e and Ø (n) have no normal components with the

exception of 1. Locate another number d with the end goal that ed =1 mod Ø(n). The plaintext is scrambled in squares, with each square having a paired esteem not exactly some number n [6].

Encryption and decoding are of the accompanying structure, for some plaintext square M and cipher text square C:

$C = M_e \bmod n$ for encryption….. (1)

$M = C_d \bmod n$ for decoding … (2)

After performing mining on the selected posts data we applied the RSA algorithm on it and both transmitter and collector must know the estimation of n. The transmitter knows the estimation of e, and just the recipient knows the estimation of d. In this manner the general population key is KU= {e, n} and the private key is KR = {d, n} Figure 2 shows the selected posts and how it securely distributed according to the mining category.

4- Evaluation metric

To evaluate the proposed method performance for secure data mining we review the popular metric used to find whether the data secured or not as TP: when predicted secure post pieces are actually named as unsecure, TN: when predicted unsecure post pieces are actually named as secure, FN: when predicted secure post pieces are actually named as unsecure and FP: when predicted unsecure post pieces are actually named as secure.

From the above we can formulate the following metrics:

$$\text{Precision}=\frac{|TP|}{|TP|+|FP|} \quad (1)$$

$$\text{Recall}=\frac{|TP|}{|TP|+|FN|} \quad (2)$$

$$\text{F1}= 2.\frac{Precision.Recall}{Precesion+Recall} \quad (3)$$

$$\text{Accuracy} =\frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|} \quad (4)$$

Where F1 used to combine precision and recall which can provide overall performance for secure mining data . for the precision , recall, F1 and accuracy the heigher is better performance [11].

5- Conclusion and Future Work:

Mining social security data is still an open field and difficult to implement since the data in sites are of different types it is also difficult to collect for business applications in data mining. A few researches have been carried out in the field of mining secure data. As described in this paper, social media data used are only the text data, which are not perfectly secured. On other hand, the information used in social media contain more than the text.

As described the performance criteria is chosen by the accuracy and F1 measure we found the performance of proposed method for categorized data as in Table 2.

Therefore, the study recommends taking image data analysis securely and sending them. Researches in the social security and social welfare data help the public and government to take important decisions about social security services.

References

[1] Roosevelt C. Mosley Jr.,(2012). "Social Media Analytics: Data Mining Applied to Insurance Twitter Posts". *FCAS, Casualty Actuarial Society E-Forum*, vol.2.

[2] Zhichao Han (2012)."Data And Text Mining of Financial Markets Using News and Social Media", school of computer science, Msc. Thesis

[3] Zoë Mullard (2010)."The Application of Social Media In The Mining Industry". The University of British Columbia, Msc. Thesis.

[4] Paul Klieber (2009)."Document Classification Through Data Mining Social Media Networks", Stetson University, Bachelor research.

[5] Sitaram Asur , Bernardo A. Huberman (2011)."Trends in Social Media: Persistence and

Decay", *the Fifth International AAAI Conference on Weblogs and Social Media*.

[6] Daniel T. Larose, A John Wiley & Sons (2005)."Data Mining Methods And Models", *Inc Publication*, pp.340.

[7] Xindong Wu and Vipin Kumar (2007). "Top 10 algorithms in data mining", *Springer-Verlag London*, vol. 14, pp.1-37.

[8] Dalia Nabeel Kamal (2010)."Proposed Enhancement algorithm for Company Employers Management using Genetic Algorithm in Data Mining". *Eng.&Tech. Journal* Vol.28, No. 2.

[9] Dr.Siddeeqy. Ameen (2005). "Security Services Provision And Enhancement In Client/Server Networks Using Aes" *IJCCCE*,VOL.5,NO.1.

[10] Dr. Soukaena Hassan Hashim (2013)."A Proposal to Detect Computer Worms (Malicious Codes) Using Data Mining Classification Algorithms" *Eng.&Tech. Journal*. Vol31,No.2.

[11] Kia Shu, Amy sliva and et al (2017). "Fake news Detection on social media A data mining perspective", cs.SI , vol3.

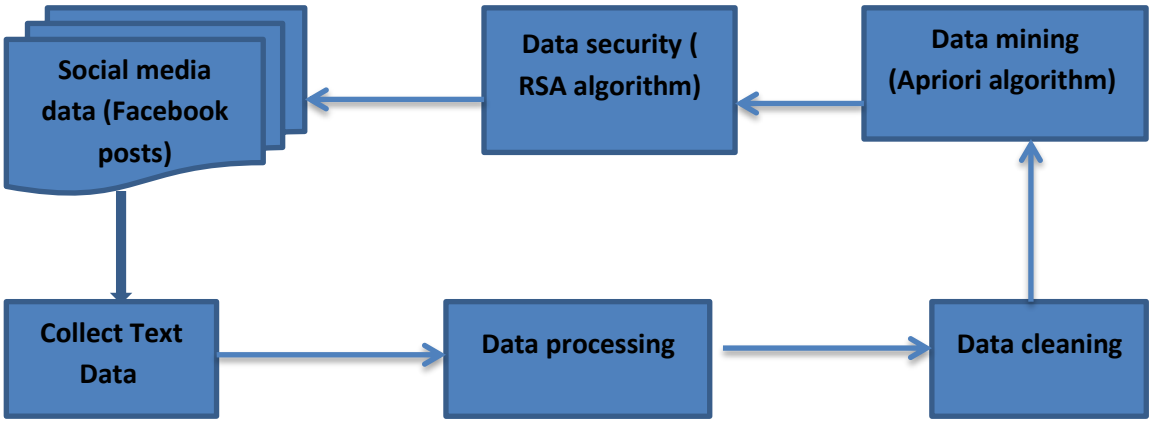

Figure1: Apriori Method of Mining Secure Social media data.

Attributes	category
User name	Facebook account
Created date	Post date
text	Text of the post
hashtag	If the post have a hashtag
tag	If the post have a tag
location	A country of the post

Table 1: Shows an example of the attributes taken from Facebook post and predefined category

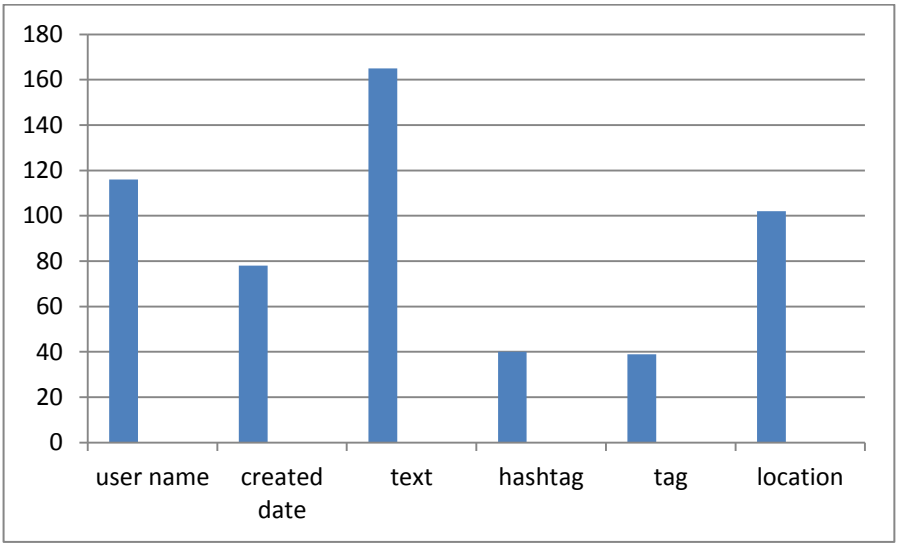

Figure 2: Shows the number of posts distributed securely according to the selected category from Apriori mining algorithm

Category	Precision	F1	Recall	Accuracy
User name	0.8514	0.4851	0.0756	0.0136
Created date	0.9165	0.5072	0.0821	0.0099
Text	0.9644	0.5039	0.0877	0.0062
Hashtag	0.9595	0.6916	0.0851	0.0077
Tag	0.9226	0.6882	0.0837	0.0088
Location	0.7432	0.8107	0.0527	0.0617

Table 2: shows the performance of the proposed method for the selected category

Algorithm Apriori
Input : D, a database of dealings
Min-Sub
Output: L, a repeated itemsets in D.
Method:
(1) L_1= find repeated 1-itemsets(D);
(2) for (k=2;$L_{k+1}\neq\emptyset$;k++) {
(3) C_k= apriori_gen(L_{k+1});
(4) for each transaction $t \in D$ { // scan D for counts
(5) C_t= subset(C_k, t); // get the subset of t that are candidates
(6) for each candidate c ∈ C_t
(7) c.count++;
(8) }
(9) L_k={ c ∈ C_t | c.count ≥ Min-Sub}
(10) }
(11) Return L= U_k L_k
Procedure Apriori gen(L_{k+1}: frequent (k+1)- itemsets)
(1) for each itemset $l_1 \in L_{k-1}$
(2) for each itemset $l_2 \in L_{k-1}$
(3) if (l_1[1]=l_2[1])^ (l_1[2]=l_2[2]) ^…^ (l_1[k-2]=l_2[k-2])^ (l_1[k-1]=l_2[k-1]) then {
(4) c= $l_1 \infty l_2$; *// join step: generate candidates*
(5) if has infrequent (c,L_{k-1}) then
(6) delete c; // prune step: remove unfruitful candidate
(7) else add c to C_k ;
(8) }
(9) return C_k;
Procedure has infrequent subset (c: candidate k-itemset; L_{k-1} frequent (k-1)-itemsets); //use prior knowledge
(1) for each (k-1)-subset s of c;
(2) if s not belong to L_{k-1} then
(3) return TRUE;
(4) return FALSE;

Figure3: Apriori Algorithm Calculation [1]

P- ISSN 1991-8941 E-ISSN 2706-6703 *Journal of University of Anbar for Pure Science (JUAPS)* Open Access
2019,13 (1) :41- 49

طريقة Apriori لتعدين البيانات الأمنة في التواصل الأجتماعي

م.م زهراء راجي محي

كلية الأمام الأعظم (رحمه الله) الجامعة، قسم الدعوة والخطابة

الخلاصة:

كما نرى من حولنا فأن مواقع التواصل الأجتماعي المتمثلة بFacebook, Youtube، Flicker, Twitter وغيرها اصبحت تشكل اهمية كبيرة في حياتنا وازدادت بالنمو خلال السنوات الأخيرة . ان هذا النمو في مواقع التواصل الاجتماعي ادى الى زيادة كمية المعلومات والبيانات المتولدة والمتداولة بين الافراد ، هذه المعلومات مهمه في اعمال الشركات والمؤسسات وحتى الأفراد، لذلك من الضروري تحليل البيانات وتصنيفها وتحديد الكلمات المفتاحية والجمل المعبرة مما يؤدي بالشركات الى ادارة اعمالها بشكل افضل مع الزبائن الحاليين والمحتملين.

وفي نفس الوقت فان بيانات التواصل الاجتماعي قد تحتوي على انواع عديدة من القرصنة والبرامج الخبيثة والاعمال الغير مرغوبة. لذا فان شبكات التواصل الاجتماعي والمجتمع والصناعة والاعمال بحاجة ماسة الى استخدام بيانات تواصل اجتماعي امنة. وفي هذا البحث قد اختيرت طريقة الApriori لتعدين وتصنيف بيانات التواصل الاجتماعي وقد اخذت Facebook كحالة لدراسة بيانات التواصل الاجتماعي بعد تصنيف البيانات وتعدينها تطبيق خوارزمية RSA الشائعة والسهلة الاستخدام لتأمين البيانات المصنفة واستخدامها بصورة مفيدة في اعمالهم.