Journal of Advances in Science and Technology
Vol. 15, Issue No. 1, March-2018, ISSN 2230-9659

Content Based Filtering and Fraud Detection on Social Networking Sites

D. P. Gadekar[1]* Dr. Y. P. Singh[2]

[1] Ph.D. Scholar, Computer Science and Engineering Department, Kalinga University, Raipur, Chhattisgarh, India

[2] Research Guide, Computer Science and Engineering Department, Kalinga University, Raipur, Chhattisgarh, India

Abstract – The online social networking (OSN) offer an extensive variety of extra data to advance standard learning algorithm, the most difficult part is separating the applicable data from arranged information. Fake conduct is indistinctly disguised both in nearby and social information, making it considerably harder to define valuable contribution for expectation models. To efficiently join interpersonal organization impacts to identify misrepresentation for the Belgian legislative standardized savings foundation, and to enhance the execution of conventional non-social extortion expectation undertakings. Finding the semantic reasonable subjects from the colossal measure of rational points from the substantial measure of client Generated Content (UGC) in online networking would encourage numerous downstream uses of shrewd processing. Subject models, as a standout amongst the most effective algorithms, have been broadly used to find the inactive semantic examples in content accumulations. In any case, one key shortcoming of point models is that they require archives with certain length to give dependable measurements adversary producing intelligent themes. In Twitter, the clients' tweets are for the most part short and loud. Perceptions of word events are immeasurable for theme models. In this research work, we proposed novel text mining based OSN fraud detection method. The proposed method is used to get fraud detection using text mining with the noise removal, bag of word algorithms, features extraction and at last naive biased for fraud classification purpose. This method helps to improve the quality in the form of result and minimize total time of processing and accuracy. The extensive experimental evaluation is presented in this work to claim the efficiency of proposed approach against the state-of-art methods using different research datasets.

Keywords-: Data, Mining, Fraud Detection, Social Networking, Security, Privacy of User.

----------♦----------

1. INTRODUCTION

It has made the world a smaller place and has opened up previously inaccessible markets of industry. The internet has delivered the large numbers of changes in the business management over the world. On other hands, denoted that multiple numbers of things regarding us, thus innovates in research paper, are analysis the data by researcher way and no scalable as private and separate as indicate them. The data mining is a parts of technology, economical and social transformation that is implemented the small world, large services impel, large interconnected, and offered unusual level of advantages. That time, the huge data is known, saved and communicate about us the private. In this chapter we will introduces an entire summary of the privacy difficulty and parameter that are utilized of data mining in business recently.

According to Berry and Line off [1] privacy is a composite difficulty, since of technology, is rising becoming a social difficulty. In the year of 2004, the Cambridge Advance Learner's Dictionary, explain the concepts of security and how to manage the personal or private data and related secret data. Presently, every formation of trading business and electronic devices, and begin that examines the private and stored for further references. These are the important problems to analysis the innovation work and privacy. We introduce that many numbers of difficulties on the security.

- Restricted are pervious on security by the social influence, and the problem is really how much data should be combined and who is manage these data.
- Each and Every customer has various objectives on privacy.
- Various levels of responsibility with view to data regarding them being possible to others.

These techniques work an import works in defined the security; protecting these security.

The data mining is a capability that represents the crucial need of business to handle their consumer relationship and calculates more easily. Data mining are mostly used in the marketing domain. Following represent the two important analysis of data mining in marketing:

- Privacy contravention may sustain legal responsibility that could output in exorbitant law suits.
- Privacy violations may output in bad press that can do analysis destruction to corporate or brand image.

In manner to understand the impacts of data mining on security of data, examines the some examples of the data mining projects in the telecom firms. The cell phone service supplies the technical ability to find out the switched cell phone location. The cell phone service supplier entire its subscribers through the sale of concurrence. The exhibitive subscriber data that are incorporate that consists of following attributes.

- Occupation.
- Age
- Income
- Banking Details

The capability of the mobile phone service supplies to innovate the mobile phone position and that why its owner, may improve the following data.

- The route typically travelled to and from work by the subscriber.
- Whether the subscriber travels during business hours, or spends most of the day in the office.
- Which shopping centres the subscriber visits over weekends or after hours.

The cell phone service supplies are implemented to use of incorporates data to locate its individual sign are large strategically, to initiated its distinct domain at the satisfactory shopping malls. Previously, the firms or industry may determine to benefits from these data by trading to another firms. For example, it is elevates to other firms who also want to be capable to position their notice more sufficiently, fast food sequence forward the advertising message to contributor and they came onto near of one of the egress. The effective projects of the above information are utilized of the data by dealer advertise on the sides of road.

Representing the income, age and business of the employee these employee investigates a specific way to increase the effectiveness of thus marketing struggle. The instance, emphasize a popular projects of data mining, it is depicted the upcoming threats that is provided security of the mining application. The significant departments of concern with regarding to information mining and security are innovated the below:

- What kind of information do you collect about your customer?
- Who is ultimately in control of that information?

The data mining formed employing data can be provide the security that the occupations results in neither of the negative influences, term, maintain valid liability or obtain the outputs of privacy violations associated with data mining attempts (New, 2004). Knowledge projects objectives at supplies data mining to marketing database for data on promising attacker (hackers), required to deficiency of analysis that was display for security problems.

The consumer may for example be informed of the reality that consolidates data about them is utilized for belling, but these are not required to enable the firm to uses the data in the mining techniques. At the end, these are the magnificent to attention to specific examination to how the information exploits in data mining was receiving starting position, and these are exploits outputs in a contravention of privacy. Privacy protection data mining is a decent research managements and statistical databases, the mining methods are examines for the sides influences maintain in data security. The main purpose of the PPDM is to enhance the methods for the changes of the native data by using the several techniques, so that personal experience and individual data endure secures after the mining scheme (Verykios, et. al., 2004). Privacy protection data mining is secured in its developments and either it will be more effective to etching entire the security in mining. Another query that, how vital mathematical mining result might be to marketing, if the private customer that the marketing try to managed it, these are cannot be obvious.

In similar time, the industry analyses the following in order to defending the firms form legal responsibility.

- Provide customers with an opt-out option whereby they have the ability to exclude themselves from being used in data mining or from being the target of directed marketing.
- Ensure that you only buy data from reputable organizations, and that the necessary permission has been obtained for making use of that data.

- Inform you customers of potential use of their information for data mining purposes, and obtain their consent prior to releasing this information to other organizations.

The data mining is techniques specialists and business expectation that finding the following accurate practices and regarding the privacy of the individual implementation better business. The poor advertising related to a person prevalence can introduce an firms region for years even when the corporations has succeeds the law and span everything this is innovates in all likelihood to because the privacy of those from who the facts changed into includes (Wang, 2003) However, both addressed a same issue, building decision tree from secret training data, the concepts of privacy are completely conflicting. One was depends on data obscuration that is, modifying the data values so real value are to expose in the year of 200 by authors Agrawal and Srikant. The second use is SMC. The SMC stands for secure multiparty computation to "encrypt" information value (Lindell and Pinkas 2000), securing that no party learn regarding other information value. The authors initially introduce SMC, then supplies aided framework on data obscuration. Author also introduces problems that have coming small analysis: How do we constrain data mining if it possible that outcome among violate privacy?

- Secure Multiparty Computation

The concept of SMC (secure multiparty computation) (Yao 1986: GIdrech, Micali, and Wigderson 1987) is that the parties involves occupy nothing but, data; they are credible third party to the whole parities imparts their input. The credible third party computed the output. SMC allow this with the credible third party. Consider the interaction between the all parties to acquire the pervious result, but this third party does not understand anything from the interaction. The computation is securely supplies to one party input and output of these are executed; we can regenerates what we seen by the third party. At that time, the regenerates mean that the distribution of the simulated view over many runs is executed identical. They may not be able to accurately regenerate each and every execution, but across time cant says the regeneration from the real execution.

2. RELATED WORK

Abraham, A., et al (2009)

In [1], authors attempt to examine a Web page as data with social aspects. Each web page is the outcomes of hiding social communication. This communication among various groups of people converts into a confirmed unification of Web page production. The external sign of this unification are the attribute of the web page that achieves the user's anticipations. Through analysis of the attributes, the authors can acquire data that can normally explain the web page. This simple explanation consist powerful data regarding the social group the page is intentional for. If the user utilized this data to clarify the search, then he detection himself as a part of a social group. For the simplification of the social feature of web page we utilized the concepts Micro Genre. In this research paper author innovated the basic theory of MicroGenre and also demonstrates practical for the identity and usage of Micro Genres.

Agrawal, R. et. al. (2000)

In [2], Author addresses the problem of privacy conservation data mining. Particularly, we examines a structure in which two parties possess secure databases want to developments a data mining techniques on the incorporate of their database, without reveal any redundant information. Our research is prompt by the necessary to both preserve privileged data and allowing it's utilized for innovation or other objectives. The above issue is a particular instance of secure multi-party execution and as thus, can be solved using called generic protocols. Data mining technique are typically ambiguous and ahead the input normally inside of huge datasets. The generic protocols in thus a case are of no experimental use and therefore more efficient protocol is necessary. The author concentrates on the problem of decision tree learning with the useful ID3 algorithm. Ours protocol is greatly efficient than generic solutions and requirement both some round of interaction and reasonable bandwidth.

Aiello et al (2002)

In [3] multiple numbers of huge graph (for example, WWW graph and Call graph) share definite universal attribute which can be report is known as "power law". In this innovation paper, we will firstly deeply study the existing research paper on power law graph. Then we will provide 4 evolution system for produce power law graphs by contributed one node/edge at a time. Author also represents the any provided edge density and wanted for in-degrees and out-degrees the output graph will are content the power law and the in/out-degree situations. In contributed, author analysis another key aspects of huge graph is known as "scale-free" in the sense that the frequency of sampling is freely of the arguments of the output of the power law graphs.

Amatriain et. al. (2010)

In [4] this innovation paper authors suggest that, to discuss an overview of the important Data Mining techniques utilized in the context of Recommender

Systems. We first explain general pre-processing approach thus as sampling or dimensionality depletion. Further, we report the large significant classification techniques, that involves SVM (Support Vector Machines) and Bayesian Networks. Author also explain the k-means clustering algorithm and describe few choice. Author also represents relation rules and that associated algorithm for an efficient training procedure. In contributed to representing these techniques, we study their application in Recommender Systems and current cases where they have been successfully enable.

Anagnostopoulos et. al. (2008)

In this innovation work author investigate that, multiple numbers of online social systems, the social connection between users are the main work in managing their behaviour. One of the strategies that can do is via social impact, the mechanisms that the activity of a user can instigate user's friends to performing in a same direction. In system where social impact exists, types of behaviour, scheme and recent techniques can spread via the network. Consequently, detecting and understanding social impact is of huge interest from all scheme point view and analysis.

This is an essential work in like manner, since there are viewpoints therefore as homophile or actually dumbfound variable that can prompt factual relationship among the movement of companions in interpersonal organization. Examination affect from these is fundamental the issue of correlation relationship from causality, a famously hard factual issue.

In this development work, creators think about this issue systematically. Additionally states fairly basic approach that copies going before wellsprings of social connection. The creator examines two simple tests that can identify affect as a wellspring of social relationship when the time arrangement of client movement is conceivable. Additionally give a hypothetical affirmation of one of the tests by demonstrate that with substantial likelihood it achieve in decision out effect in rather regular method of social relationship. We replicate our own tests on a different quantities of illustrations conspire by anyplace creating action of hubs on a real system comparing to one couple of approach. Ultimately, we apply then to genuine tagging data on Flickr, uncover that while there is gainful social relationship in tagging conduct on this framework, this connection can't be credited to social impact.

Bulkley et al (2006)

In this innovation study of hypotheses regarding, the systematic and important user of social networks by the specific group of white collar workers. Also study of existing concept that related to network scheme to output and put sending two recent hypotheses. The first extension combined analyzes hypotheses with social networks; suggest that optimal network attributes spread across the course of a career form these favouring survey to these favouring survey of information and association. The second concerns efficient activity of data via a network, innovates that rapid short interaction outperforms infrequent lengthy interaction. By using a exclusive data sets consists email scheme and accounting report for some dozen executive recruiters, we investigate statistically important comparison related to network (1) scheme (structure) (2) flow (3) age. Consistent with existing concepts, large central position is related with larger output. Consistent with the two innovation hypotheses, analysis scheme between early career recruiters and profiteering scheme between senior recruiters are both positively related with performance, while large rapid shorter message are related with larger output. Output of this innovation have the potential to generate a fully understanding of various mode of efficiency related with social networks.

Cai, D et al (2005)

In this innovation paper author investigate that, Social network analysis has mesmerize much focus in recent years. The community mining is the principle approach to removing data in social network examination. The majority of the current procedures on group mining assume that there are just single kinds of relationship in the system, and further, the mining yield is autonomous of the customer's important or enjoying. Despite the fact that in real, there introduce different, heterogeneous interpersonal organizations, these are exhibit the specific sorts of affiliation, and these kind of connection are principle part in particular work. Along these lines mining framework by close just single kind of connection may lose a considerable measure of beneficial concealed group information and may not be versatile to the diverse information fundamental from special clients. In this exploration paper, we methodically investigate the issue of mining concealed groups on various social networks.

Our strategy to SN analysis and social mining display an objective modification to technique from the traditional one, the modification from user freely analysis to multi-network, user-dependant, personal network and query based analysis. The practical outcomes on Iris Group and DBLP data set current the yielding of our innovation model.

Caruana, R. & Niculescu-Mizil, A. (2004).

In this innovation author investigate that, multiple numbers of tests can be utilized to analysis the performance of supervised learning. Unique criteria are important in different setting, and it isn't generally essentially which criteria to utilize. The further trouble is that learning strategies that activity well on one standard may not activity great on another criteria. For instance, the help vector machine (SVM) and

boosting are plan to improve rightness, though neural nets routinely investigation cross entropy or squared mistake. We sort out a functional examination an alternate learning calculation (neural nets, packed away and helped trees, SVM and supported stumps) to separates 9 Boolean grouping execution measurements: F-Score, Accuracy, and Area under the ROC Curve, Precision/Recall Break-Even Point, Lift, Cross Entropy, Average Precision, Probability Calibration and Squared Error. The MDS remains for Multidimensional scaling. The MDS exhibits that measurements traverse a little Dimensional complex.

The three measurements those are reasonable when supposition is clarified as probabilities: adjustment, cross entropy and squared mistake, lay in one segments of metric space far from measurements that in light of neighbouring way of the likely esteems: make back the initial investment point, normal exactness, lift and ROC region. In the middle of them drops two measurements that in light of separates expectations to an edge: accuracy and F-score. An envision, huge edge strategies along these lines as SVMs and helped tress have better effectiveness on measurements like rightness. In any case, action ineffectively on likelihood like as squared mistake. What was not expects the edge approach have better effectiveness on requesting measurements in this manner as normal exactness and ROC territory. We speak to a current metric, SAR that coordinates rightness, squared mistake and ROC locale in to one metric. relationship investigation and MDS speak to that SAR is centre finds and consolidates great with another measurements, speaking to that it is a superior regular goals metric to utilize when huge specific criteria are not known.

Dwyer et. al. (2007).

In this innovation work author investigates that, it is not well understood how privacy cover and trust impact social communication within social networking sites. An online review of two renowned interpersonal interaction destinations, MySpace and facebook separates learning of trust and security worry, alongside availability to share information and actualize late affiliation. Individuals from all site recorded same levels of security concern. The facebook individuals show advantageously more noteworthy trust in facebook and its individuals, and were more status to share identifying information. All things considered, MySpace individuals recorded vitally more event utilizing the site to achieve new human. These yield demonstrated that in online correspondence, trust isn't as required in the developing of late relationship as it is in vis-à-vis experiences. The creator likewise demonstrate that in an online website, the current of trust and the eagerness to share information don't consequently changed over in to late social correspondence. The examination speak to online connections can executes in locales where perceived trust and protection shields are frail. The writing questions were producing to get impression of put stock, by and large utilization of the webpage, web protection concern, advancement of late affiliation and data sharing. These questions created from a qualitative examination organize by author Dwyer in the year of 2007. Each and every issue was re-worded for the two literatures. For e.g. the facebook observation involves the issue “I try to send a message to a friend using Facebook messenger rather than through using email.” The MySpace versions suggest “I prefer to send a message to a friend using MySpace rather than through using email.”

Gross, R., & Acquisti, A. (2005).

In this research paper author investigate that, Take a part in social networking sites has seriously rising in present year. Services are Tribe, Friendster and facebook enable lots of private to create online profiles and forwarding users personal data with too big network of relatives and friend. In this innovation paper author examine patterns of data disclosure in online social networks and their security association. They analysis is one of the behaviour of more than 4,000 Carnegie Mellon University students who have merge a famous social networking site provision to colleges. They calculate the large amount of data they reveal and study their usage of the site privacy settings. They overview conventional attack on different aspects of their security, and they represent that only minimum percentage of users alter the maximum pervious security preferences.

3. RESEARCH METHODOLOGY

3.1 Problem statement

The following problem statement briefly defines the boundaries and environment of this project: The Development of an android application which has the capability of using the concepts of augmented reality to submerge the virtual information of user’ surroundings by detecting and tracking user’s location in real time. As the android GPS is notified, the application is fully location aware which keeps the track of user’s location. When the user points the camera in a specific direction, the application tracks the camera orientation and displays the records of a specific place. Then the application keeps on updating the information as the direction changes. The additional information is displayed with the help of "Google" databases. The information when gathered is then displayed to the live feed of camera which helps the users to interact in a more reliable

way. Option for viewing the places in map view with the help of Google Maps is also available.

3.2 System architecture

The architecture is help to the OSN services is a three-tier strategy in figure. 1. The first layer of architecture is SNM (Social Network Manager), generally main goal to supply the fundamental functions are profile and relationships, since the second level aided the SNAs. The helps SNAs may in revolve needs a contributed layer for their necessary GUI. According to this liking scheme, the investigation model discus in the previous two layers. In particular, user interaction the

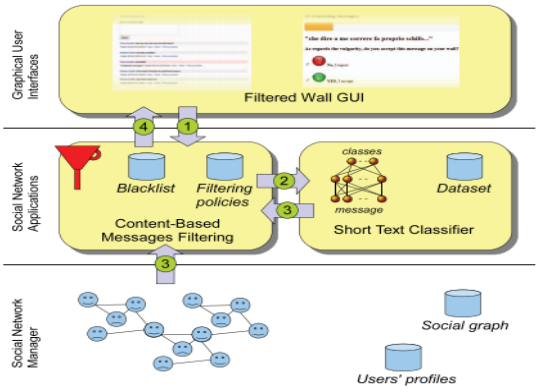

Figure.1. Filtered wall theoretical scheme and the direction messages follow, from writing to communication

3.3 Algorithms

Algorithm 1: Noise Removal Algorithm

Input: P here P is Document file

Step 1: d1 = EscapingHtmlCharacters (P)

Step 2: d2 = DecodingData (d1)

Step 3: d3= Apostrophe_Lookup (d2)

Step 4: d4= RemovalOfStopWords(d3)

Step 5: d5= Apostrophe_Lookup (d2)

Step 6: d6=RemovalPunctuations(d5)

Step 7: d7=RemovalExpressions(d6)

Step 8: d8=SplitAttachedWords(d7)

Step 9: d9=Slangslookup(d8)

Step 10: d10= StandardizingWords (d9)

Step 11: d11= RemovalOf Url (d10)

Step 12: Stop

Step 1: Breaking out HTML characters: Data acquire from the web generally that involves a larger contains of the html data thus are & < and > this obtained combined in the actual data. It is utilized to obtained free of these objects. One approach is to explicitly extracted from the data contain by using of particular regular expressions. The other model is to utilized suitable packages and modules; these are transferring the objects to the html tags. For instance: < is transfer to “<” and & is transfer to “&”.

Snippet:

import HTML Parser

html_parser = HTML Parser. HTML Parser()

tweet = html_parser.unescape (original_tweet)

Output:

“I luv my <3 iphone& you’re awsm apple. DisplayIsAwesome, sooohappppppy http://www.apple.com”

Step 2: Decoding data: It is a procedure to converting data from the ambiguous symbols to general and easy to learn character. The Text information are divergent forms of decoding for example “UTF8” and “Latin” etc. Therefore, for evaluation, it is required to maintain the fully data in standard encoding pattern. UTF-8 encoding is broadly confirmed and is supported to use.

Snippet:

tweet = original_tweet.decode("utf8").encode(‘ascii’,’ignore’)

Output:

“I luv my <3 iphone& you’re awsm apple. DisplayIsAwesome, sooohappppppy http://www.apple.com”

Step 3: Apostrophe Lookup: To avert any word impression complex in text, it is suggested to manage correctly scheme in is and to stand for by the rules of CFG. When digression are utilized, a possibility of disambiguation rising.

Snippet:

APPOSTOPHES = {“'s" : " is", "'re" : " are", ...} ## Need a huge dictionary

words = tweet.split()

reformed = [APPOSTOPHES[word] if word in APPOSTOPHES else word for word in words]

reformed = " ".join(reformed)

Outcome:

"I luv my <3 iphone& you are awsm apple. DisplayIsAwesome, sooohappppppy http://www.apple.com"

Step 4: Removal of Stop-words: When data evaluation necessary to be data compelled at the world level, generally obtained world should be eliminated. One can either developed a larger set of stop-world or one can user predefined language specific libraries.

Step 5: Removal of Punctuations: The entire punctuation denoted allow to the preference to be deal with. For instance: ".", ",","?" are the main punctuations that should be received during others necessary to be eliminated.

Step 6: Removal of Expressions: Textual data may contain people expressions like [Crying], [laughing] and [Audience paused]. These expressions are generally non-applicable to data content of the speech and since they necessary to eliminated. The general regular expression can be used in this situation.

Step 7: Split Attached Words: We people in the social conference produces the text data, which is fully non-formal in nature. The large numbers of the tweets are attends with lots join word thus Playing In The Cold and Rainy Day etc. these object can be separates into their normal forms by using simple rules and regex.

Snippet:

cleaned = " ".join(re.findall('[A-Z][^A-Z]*', original_tweet))

Outcome:

"I luv my <3 iphone& you are awsm apple. Display Is Awesome, sooohappppppy http://www.apple.com"

Step 8: Slangs lookup: Repeated, the social media contain of consent of dialects world. The world should be converted into standard worlds to created free text. The words like luv are changed or transfer to love, Helo to Hello. The same model of apostrophe sees that are utilized to transfer slangs to standard words. A numbers of sources are possible on the web, these are supplies sets of all available slangs, this would be your holy chalice and you could use them as lookup language for transformation use.

Snippet:

tweet = _slang_loopup(tweet)

Outcome:

"I love my <3 iphone& you are awesome apple. Display Is Awesome, sooohappppppy http://www.apple.com"

Step 9: Standardizing words: consistently world are not in correct formats. For instance: "I looooveee you" converted in to "I love you". General rules and regular expressions can be assists solve these cases.

Snippet:

tweet = ''.join(''.join(s)[:2] for _, s in itertools.groupby(tweet))

Outcome:

"I love my <3 iphone& you are awesome apple. Display Is Awesome, so happy http://www.apple.com"

Step 10: Removal of URLs: the hyperlinks and uniform resources locater in the texta content thus are review, tweet, post and comments are eliminated.

Final cleaned tweet:

"I love my iphone& you are awesome apple. Display Is Awesome, so happy!",<3 ,

Algorithm 2: Keyword extraction

Step1. Split the document into an array of words, breaking it at word delimiters (like spaces and punctuation).

Step2. Split the words into sequences of contiguous words, breaking each sequence at a stop word. Each sequence is now a "candidate keyword".

Step3. Calculate "score" of each individual word in the list of candidate keywords.

Step4. For each candidate keyword, add the word scores of its constituent words to find the candidate keyword score.

Given an input document, on which we want to extract keywords,

1. Split the document into an array of words, breaking it at word delimiters (like spaces and punctuation).

2. Split the words into sequences of contiguous words, breaking each sequence at a stop word. Each sequence is now a “candidate keyword”.

Consider the short text “A scoop of ice cream.” We break this into words to get

["A", "scoop", "of", "ice", "cream"]

Step 2 arranges these words into sequences by avoiding stop words. The words “A” and “of” are bound to be on any stoplist you’re using. So, by reading the array from left to right, skipping stop words, and creating a new candidate keyword every time a stop word is encountered, we obtain two candidate keywords:

["Scoop", "ice cream"]

3. Calculate the “score” of each individual word in the list of candidate keywords.

This is calculated using the metric:

Degree (word)/frequency (word)

It’s easy to understand what the frequency of a word is. It’s simply the most time of the word obtained in the overall list of candidate keywords. So our word frequencies are:

frequency("scoop") = 1

frequency("ice") = 1

frequency("cream") = 1

But what is the degree of a word, and more importantly what does it mean? The explanation involves a bit of (simple) graph theory. If you don’t want to get into that, you can skip the next paragraph; the intuition behind the degree of a word is summarized after the explanation. The degree of a word in this context is similar to the degree of a node in a graph. Let’s represent this problem as a graph. Draw an undirected graph with each content word as a node. Connect two nodes together if they appear in the same candidate keyword. The more connections a node has (i.e the higher its degree), it tends to mean that the word occurs often and in longer candidate keywords. So, the degree of a word represents how frequently it co-occurs with other words in the candidate keywords. The degrees of the words in our sample sentence are:

degree("scoop") = 1

degree("ice") = 2

degree("cream") = 2

If that didn’t make sense, to find the degree of word W, all you need to do is count the number of words that obtained in user keywords containing W, including W itself. Consider the following list of candidate keywords, taken from an example that is used in the original RAKE research paper:

Compatibility – systems – linear constraints – set – natural numbers –

Criteria – compatibility – system – linear Diophantine equations –

Strict in equations – non strict in equations – Upper bounds –

components – minimal set – solutions – algorithms –

minimal generating sets – solutions – systems – criteria –

corresponding algorithms – constructing – minimal supporting set –

solving – systems – systems

If we want to find the degree of the word “set”, degree(“set”), we simply count the total number of words that appear in candidate keywords containing the word “set”. So, degree("set") = 6 (see the bold candidate keywords)

degree("natural") = 2 (see the underlined candidate keyword)

Degree(“set”) > degree(“natural”) because “set” co-occurs with other words more frequently than “natural”. Also, if there were a long candidate keyword containing, say, 5 words including some word W, then the degree of W would be at least 5. Thus, a higher word degree could also indicate that a word appears in a long candidate keyword. We now know what the degree and frequency of a word are, but what about their ratio, which is the word score metric used in RAKE?

word_score = degree(word)/frequency(word) = ???

The word score has the word degree in the numerator and the word frequency in the denominator. This means that the word score is proportional to the word degree and inversely proportional to the frequency. We know that the degree is high when a word appears frequently, especially with other words, and when the word appears in long candidates. The frequency is high when a word appears frequently, regardless of where it appears. So, the RAKE word score metric disfavours words that appear too frequently and not in long candidates, and favours

words that primarily obtained in longer candidate keywords.

4. For each candidate keyword, add the word scores of its constituent words to find the candidate keyword score.

Take the first one-third highest scoring candidates from the list of candidates as the final list of extracted keywords.

In our first example, the word scores are:

word_score = degree(word)/frequency(word)

word_score("scoop") = 1/1 = 1

word_score("ice") = 2/1 = 2

word_score("cream") = 2/1 = 2

score("scoop") = word_score("scoop") = 1

Algorithm 3: Bag of words

Step 1: Collection of Data

Step 2: Design the Vocabulary

Step 3: Create Document Vectors

Step 1: Collect Data

Below is a piece of the first some lines of text from the book “A Tale of Two Cities” by Charles Dickens, taken from Project Gutenberg.

It was the best of times,

It was the worst of times,

It was the age of wisdom,

It was the age of foolishness,

For this littlie example, let’s consider each line as a different “document” and the 4 lines as our entire corpus of documents.

Step 2: Design the Vocabulary

Now we can make a list of all of the words in our system vocabulary.

The unique words here (ignoring case and punctuation) are:

- “it”
- “was”
- “the”
- “best”
- “of”
- “times”
- “worst”
- “age”
- “wisdom”
- “foolishness”

That is a vocabulary of 10 words from a corpus containing 24 words.

Step 3: Create Document Vectors

In this step we calculate score the words in each document. The aim is to turn each and every document of free text into a vector that we can use as input or output for a machine learning model. Because we know the vocabulary has 10 words, we can utilize a constant-length document representation of 10, with one position in the vector to score each word. To mark the presence of words as a Boolean value the simplest scoring method, 0 for absent, 1 for present Using the arbitrary ordering of words listed above in our vocabulary, we can go through the first document (“It was the best of times“) and change it into a binary vector. The scoring of the document would look as follows:

- “it” = 1
- “was” = 1
- “the” = 1
- “best” = 1
- “of” = 1
- “times” = 1
- “worst” = 0
- “age” = 0
- “wisdom” = 0
- “foolishness” = 0

As a binary vector, this would look as follows:

[1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

The other three documents would look as follows:

"It was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]

"It was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]

"It was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]

All series of the words is regularly not helpful and we have a fixed directions of discover features from any types of files in our entity are disposed for utilize in modelling. The recently produce files that overlap with the language of known words, but it not assured about words from of the language, can be quiet be encoded, where only the existence of known word are save and the unknown word are eliminating.

4. RESULTS AND DISCUSSION

In this section we discussed about the comparative analysis and results of the research as follow

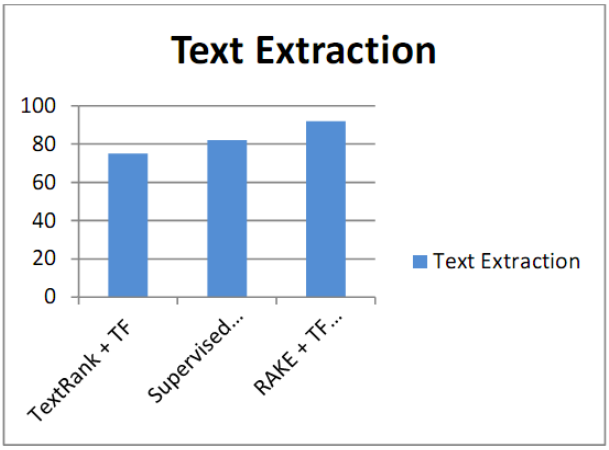

Figure 2. Text Performance Bar Graph

This is comparison of text performance of both existing Text Rank + TF, Supervised Learning +TF and proposed system. The performance of proposed method (RAKE + TF) is better than existing methods as we are combining different techniques line data cleaning, feature extraction etc.

Below graph shows comparison between Text Rank + TF, Supervised Learning +TF and proposed method of feature extraction. The feature extraction is more in case for our proposed method.

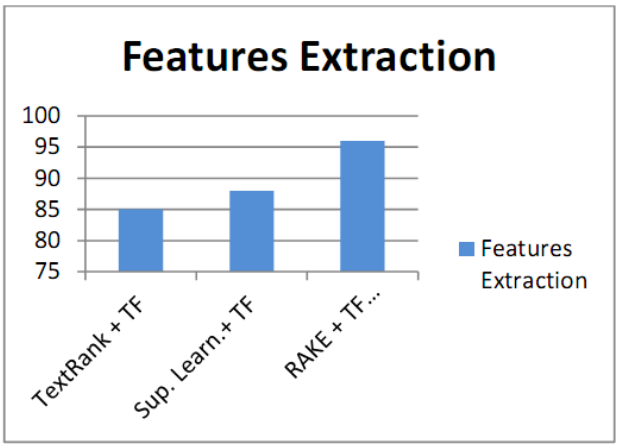

Figure 3. Feature Extraction Bar Graph

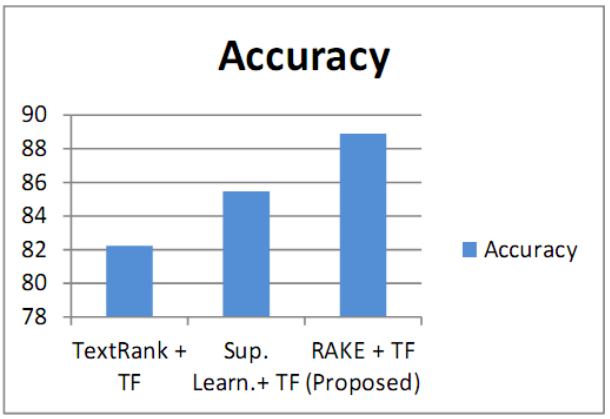

Figure 4. Accuracy Graph

5. CONCLUSION AND FUTURE WORK

This system is developed to filter redundant message from OSN walls. The wall that controlled the redundant message (can be considered as fraud) called as the FW (Filtered Wall). This system extends to denote the thought concerning the filtering system. In contributed, this system extend to survey and analysis the approach and techniques restricting the interferences that user will do on the required to filtering protocols with the objectives of going the filtering model and therefore the variable rules of existing based model. This system extends to allow the text filtering algorithm at pre-processing phase thus on reason the message and believe computation for message. In this innovation, proposed system builds in a manner in which each and every amp and message supplies the trust goodness of user.

The project has implemented almost entire requirements. Moreover, requirements and enhancements that requirement can simply be complete since the coding is mainly scheme or modular in form.

It has represented a system to filter out redundant content such as text messages from OSN walls. To compulsory customizable content-based filtering rules

the system utilizes a soft classifier. Furthermore, the scalability of the model in theory of filtering selection is good through the industry of BLs. It is needed to provide security to blacklist management system and Filter rules. This research innovation system have overcome the drawback of existing system to instead of blocking user to notify message to that user by using mail. For future, image can be filter in online social network by using OCR.

REFERENCES

Abraham, A., et. al. (2009). Social aspects of web page contents. In: Abraham, A., Sn´asel, V.,Wegrzyn-Wolska, K. (eds.) Proceedings of the International Conference on Computational Aspects of Social Networks, CASoN 2009, Fontainebleau, France, 24–27 June 2009, pp. 80–87. IEEE Computer Society, Washington, DC (2009)

Agrawal, R. and Srikant, R. (2000a). Privacy-Preserving data mining.In Proceedings of the ACM-SIGMOD International Conference on Management of Data (SIGMOD). pp. 439–450.

Ahuja, R. K., Magnanti, Thomas L., and Orlin, J. B. (1993). Network Flows: Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs, N.J.

Berry M.J.A. and Linoff G.S. (2000). Mastering Data mining: The art and science of customer relationship management, Canada Wiley.

Blazevic L. Agrawal, R.; and Srikant, R. (2000). Privacy-Preserving Data Mining. In Proceedings of the ACM SIGMOD International Conference of Data, pp. 439-450.

Evfimievski, S. (2002). Randomization Techniques for Privacy Preservation Association rule mining. SIGKDD Explorations 4(2); pp. 43-48. 1.

Goldreich, O.; Micali, S; and Wigdeerson, A. (1987). How to Play any mental Game. In Proceedings of the Nineteenth annual ACM Symposium on the theory of computing, pp. 218-299.

Lindell, Y.: and Pikas, B. (2000). Privacy Preservation Data Mining. In Advances in Cryptology-CRYPTO 2000.

New W. (2004). Pentagon failed to study privacy issues in data mining effort, IG says.

Verykios, V.S.; Bertino, E.; Fovino, I.N.; Provenza, L.P.; Saygin, and Theodoridis, Y. (2004). State-of –the-art in Privacy Preserving Data Mining. SIGMOD Record. Volume 33, Issue 1: pp. 50-57.

Wang J. (2003). Data Mining Challenges and Opportunities. London, IRM Press.

Corresponding Author

D. P. Gadekar*

Ph.D. Scholar, Computer Science and Engineering Department, Kalinga University, Raipur, Chhattisgarh, India