A Hybrid Approach to predict negative comments on wall using Sentiment Analysis

M. Gunasekar, S. Amruthavarshini, L. Hema Malini, T. Santhi

Abstract: In modern life, the world is trending towards the Internet. Now-a-days Internet leads Human Beings. In last few years, Internet acts as the means of communication in order to share our views, ideas, Information with others. This is achieved by social networking sites. Such interchanges might include drivers sorts of substance such as text, image, audio and video data. The Security measures provided by Internet are efficient but it is limited. While Socializing one can access the profile of another which are involved in wall and one can share data such as image, text, video etc. The major issue in user wall is the inability to control the information Such as comment, tag, memes that is posted on their own wall. To address this issue, we proposed a system which controls the text contents posted on the individual's wall. The system allows user to customize the polarity filtering criteria on the walls and text mining based is based on Short Text Classification(STC). STC is an automated classifier which label comments as positive or negative. The friends who have been posting negative comments continuously can be blocked with a real time mobile intimation to the owner of the account.

I. INTRODUCTION

Data Mining is the computing process in which it discovers patterns for data sets in huge involving intersection methods of machine learning, statistics of machine learning and database Systems. In other word Data Mining is the process of extracting datafrom unprocessedinformation list and converting it to an unambiguous format for further use. It is alsodefined as the analysis of the "knowledge discovery in databases"(KDD) is the field at which the coincide of computer science(CS) and statistics take place. It uses methods at the intersection of artificial intelligence, machine learning, statistics of machine learning, and systems. Apart from the unprocessed data evaluation,it alsopertain to database and data management of the information, pre-processing the data model and consideration of interface, interest ofMetrics, complications, consideration, post-processing of the invented structures, visualization, and online refurbish of information's in database.

The task of data mining involves the examinethe huge sum of data to be carried out formerly undetermined, patterns such as groups of atypical records for irregular detection, seeking for association rule mining, sequential pattern mining and records for cluster analysis. The task is carried out by database techniques like spatial indices. Then the patterns are

seemed to be different kind of abstract of information, and it is carried for further analysis. For example, it is done in machine learning as well as predictive analytics. The data mining procedure identifies N-number of cluster of data which can be used to attain appropriate prediction results of forecastby a decision support system. These are additional steps in KDD process.

Data Mining is the process in which it poses queries in a huge amount of Data Sources and also extracting patterns and trends using Statistical and Machine Learning Techniques. It also integrates various technological methodology including database management, statistics and machine learning. Data mining has applications in numerous disciplines including medical, financial, defence and intelligence. Data mining tasks include clustering, classification, anomaly detection and making association. For example, data mining can extract various terms such as people, places or words. Now-a-days there are many developments in data mining. The process is to extract data to discover hidden connections and predictableupcoming that causes extended period of time. Occasionallyit is ensured to as "knowledge discovery in databases,(KDD)". The phrase "data mining" was coined after 1990s.The basisof data mining consists of theirscientific disciplines such as statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions). The data mining technology keeps move forward to keep phase with unlimitedpossibility of huge data and provide computing power. Various data mining techniques have been developed.

These techniques include neural networks, extracting associations, decision trees, inductive logic programming, fuzzy logic and rough sets. Furthermore, data mining has gone beyond mining relational databases to mining text and multimedia data. It is also being applied to information security and intrusiondetection. While there have been many practical developments, we still have major challenges. One of the most important challenges is scalability. If data mining is to be useful we need to mine very large databases. Therefore, it is critical that we need to understand the limitations of the data mining algorithms. To understand the limitations, we need to study the foundations of data mining. We need to explore the time and space complexity of the algorithms. There are techniques such as rough sets that have underpinnings in logic and mathematics. One needs to explore these techniques for data mining and examine the computational complexity aspects. We also need to understand the complexity of the various search algorithms being used for market basket analysis.

II. RELATED WORK

The prime requirement of data mining are Sources of data, pattern evaluation module,data mining engine, graphical user interface and knowledge base

a) Sources of data:

The actual source of data is Database of a data, server of data warehouse, text files, World Wide Web and other related documents. To be successful, you must have huge amount of ancestral data for data mining. The data's are stored in databases or data warehouses by organization. Data warehouses may contain spreadsheets, databases, text files or other kinds of detailed repository.The Statistical information also resided in plain text or spreadsheets. Internet or the World Wide Web is huge origin of data.

### *Different Processing techniques:*

The data will be passed to the database or data warehouse server after cleaned integrated and selected. In spite of the data is from non-identical origin and in unlike, unusedformat exactly indata mining process because the data are unfinished and reliable. So, the data must be cleared and integrated initially.Then many data are necessary and it must be accumulated from different data sources and only interested data's are needs to be selected and transformed to the server. These processes are too ridiculous. Many techniques may process on the data as part of cleaning, integration and selection.

b) Database or Data Warehouse Server

The stored data to be processed is contained in database or data warehouse server

c) Data Mining Engine:

The data mining engine in which it has number of modules that is used for carrying out the data mining mission including classification, features,association,forecast, clustering and time-period examine.
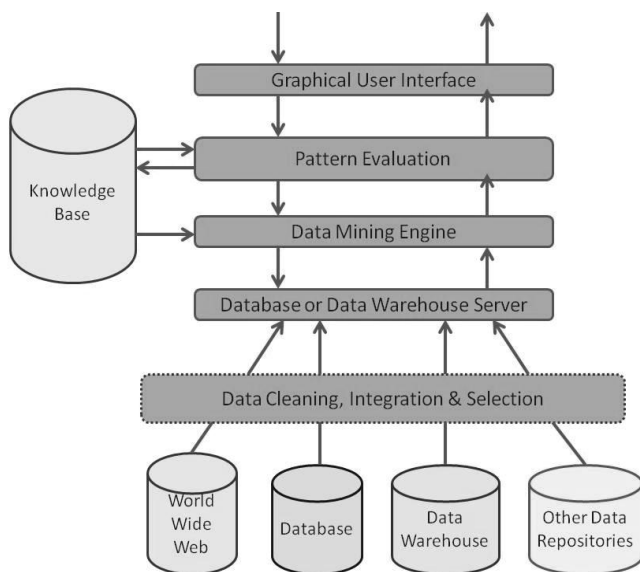
Fig 1: Architecture of Data Mining

d) Pattern Evaluation Modules:

The pattern evaluation module is the one which is responsible for measuring the pattern by using a starting point value. In other words, it also interconnects with the mining of

data's in engine that is to be focused with the identifications towards compulsive patterns.

e) Graphical User Interface:

The Graphical User Interface(GUI) acts as an interfaceto the end user and mining of data's in the system. Without knowing the actual complexity of the process, this module assists the one to use the system easily. When the user is supposed to specify a query or a mission, then the module will communicate with the data mining system. Then the outcome will be displayed in unambiguous aspects.

f) Knowledge Base:

The knowledge base is the one which is used for the search guidelines or analysing the interested outcome patterns. It consists of the data from the user experience and user belief that will be helpful in the process of mining the data. The data mining engine get input data from the knowledge base in order to make the outcome well founded as well as accurate. The knowledge based as well as pattern evaluation module conjointon a frequent basis to get input data's as well as refurbish it. "Topic and role discovery in social networks"-2005[3]

III. SENTIMENT EXTRACTION AND CHANGE DETECTION

The work presents the extraction of information about the user'ssentiments whether it is positive/neutral/negative that has been declared by the user.On the other hand, there is investigation of detection of sentimental changes that are being detected according to the "usual" sentiment of each user. We present the methods developed with each purpose below.

3.1 Preprocessing

The process of all words that are being converted into the lowercase later it huntouts idioms and join the words. For example,"de male fe" that is meant to be duplicityas it is converted into "de mala fe", afterwards functionally process as a ideal work

3.2 Segmentation of sentences

The comment or messages are differentiated into sentence by using the punctuation marks that to be considered as separators such as commas or semicolons that are the part of emotions.

3.3 Partial Tokenization I

Here, tokenization done can be done extracting each sentence using white spaces only.

3.4 Detecting the emotions

Here, emotions can be detected by using the classifier scratches the text and it gets occupied in two file based on text that contain affirmative and pessimistic emotions. The lists of emotions that it included in thefiles contain that excerpt from Wikipedia along with some other collection of additional data's.

3.5 Complete Tokenization

Here, punctuationmarks such assemicolon are consideredto be asseparators that vow to obtain the resultant set of tokens for each and everysentence.
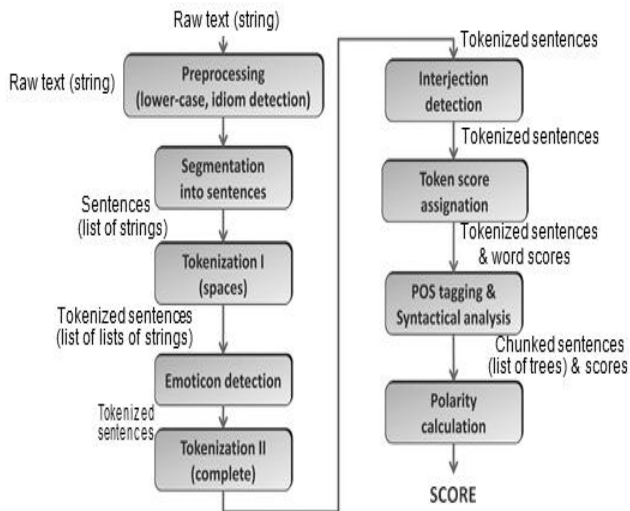

Fig 2: Representation of sentiment Extraction

3.6 Interjection Detection

The step hasidentifying and describing interjections. Those who express feelings, such as "jejeje" or "jajaja", are represented in affirmative signs whereas interjections such as "joopeee"(distress) or "uff" (tired) are represented in pessimistic signs. The implementation is detected through regular expressions, at the maximum. The interjections are extendedby set of inputs containing its ownword or repeated letters. For eg: "jope" means distress whereas "jjjooooooooopeeeee" represents a stronger (distress) sentiment. Were frequent expressions are not used in the step; many interjections would not be identified e.g.., "jajajaja", which relates the frequent emotions "ja+".

3.7 Token Score assignation

This phase is about the score that are assigned to each token;if 1pass on a affirmative comment or sentiments, if 0 is fair, and if 1is pessimistic, to assign a score the classifier is about to check whether the token are theaffirmative or pessimistic emotions, a affirmative or pessimistic interjection, else it matches to the words that are reserved in the sentiment lexicon (L).

3.7.1 Building the lexicon

The sentiment lexicon was built earlier starting from the Spanish Linguistic Inquiry and Word Count(LIWC) (Ramirez-Esparza, Pennebaker, Garcia and Suria,2007).The classification falls on either of these two categories,

- Positive: Affirmative emotions, Affirmative sentiment, optimistic.
- Negative: Pessimistic emotions, anger, sadness, death to swear.

3.7.2 Eliminating repetitive letters

Here it is used to checkwhether it has appeared in any of two dictionaries category either is affirmative or pessimistic, respectively.The data that are in Facebook are generally represented in a common language. It is repeatedlyused to identify words using the frequent letters (e.g.

"fenomenaaaaal"-"great" in English)or characters of non-alphabetic for each tokens, the messages that are appeared are not found in the dictionary form, the letters that are occurring twice in a row are reimburse only by single processing as well as the unique words are looked for in the dictionary. In case the word not found in dictionary then it is lessenedto its lexeme.

3.7.3 Spell Check

Here, the word that is not reliable with any word then it is analyzed using the spell check. Since Facebook messages consists of incorrect spellings, in order to eradicate the spelling checker incorporated to classifier: GNU Aspell, included most of the Linux distributions. Although, the spell checkis supposed to be implemented carefully. As some corrections are suggested, provide worst result in classifier. For example, for the word ''Dani'' (a popular Spanish nickname) the suggested correction was ''daño'' (damage, a pessimistic word). With the purpose of avoiding these situations, a set of words inlistshould be uncorrected, including names and surnames, with in the dictionary as being created and incorporated.

3.8 Polarity Calculation

Polarity calculation is calculated by using number of tokens which is used to convey sentiments based on their grammatical classification(i.e. noun, adjective, interjection or verb) that are calculated. The words that appears repeatedly in the text form (i.e. determinant, prepositions, etc.), are not taken into an account, as they are "stop words"for a sentiment analysis. Once the token is scored, polarity score of a sentence is calculated. The sum of the scores is divided by the sum of all the candidates to receive a score, the score that is obtained from 1 to +1 "Viral Misinformation" – 2015[8].

IV. PROPOSED SYSTEM & RESULTS

An increasing variety of individuals is increasingly approaching to the social networking sites. Social media is changing into a vital topic for research in several fields. As variety of individual's using social network are growing day by day, to speak with their peers in order that they'll share their personal feeling every day and views are created on massive scale. In nowadays several firms are using social media selling to advertise their merchandise or brands, thus it becomes essential for them that they'll be ready to calculate the success and utility of every product.

Sentiment analysis is that the Natural Language Processing (NLP) task handling the detection and classification of sentiments in texts. The early work that is referred to be original domain that includes filtering of information, successive papers have represented as various domains including newswire articles, This is carried out by an electronic mail.Complex filtering systems that consists of multi-labelled text classification, automatically labels messages into partial thematic section.

Content-based filtering is based on the usage of the Machine Learning (ML) concept in whichthe classifier itself induced to learning from a set of pre-classified examples. The related work that differs in adopted feature extraction method and model learning. The extraction process characteristic is

mapped to the text with a compact representation of the content. It constantly applied to training and generalization phases. Many research proved that Bag of Words (BOW) is the approach that yields large amount of sophisticated text identity that may have superior semantics but inferior statistical quality. Classification is the one which separates or orders the objects into classes. There are two phases in classification algorithm: The algorithm which findsthe model for the class attributes as a function of other variables of the datasets then it is applied for determining related classes for each record. The predefined categories are assigned into text by the text classification.

Text categorization depends on the technique which retrieve informationsuch as indexing, inductive construction of classifiers and evaluation technique. By using machine learning the classifier can be able to study how to make the classification of documents based on the features extraction of training data. The unstructured data is used for social content mining such as text. Mining of unstructured data that had hidden information and the information extracted from different text sources from previously unknown information.

Social content mining requires the data mining application of text mining application of data mining techniques.The Social Network is trending to be the most worldwide popularconjoint medium. In order to share communicate, views as well as thoughts based on human life. The major issue in user's wall is the inability of the user to control messages or comments posted on their own private space. In order to obtain a blunt carry out on the messages post on the wall we proposed a system allowing Online Social Networking (OSN) system has an authority to control the messages. This is achieved by polarity filtering basis to be applied on the walls and the classifiers automatically label the messages with the help of content based filters.

Machine Learning(ML) is used as text categorization technique that automatically assignsthe short text messages with specified content that is based on the classifications. The building of short text classifier (STC) concentrates in the process of extracting as well as selecting the set of characteristics and discriminating features. Here a database with the characterized word is built and used to check the words. If the wall consists of any negative word such as indecent word, vulgar words then they will be sent to the black list to filter at those words from the messages. Finally the message with the positive words will be posted on the user's wall.Thesystem automatically filters the unnecessary information on basis of information content as well as information creator relationship and features.

Social network app:

- Social network refers to the interaction among people in which they communicates their views share, and exchange information and ideas in virtual communities and networks
- GUI is the design which is the type of user interface that provides were to interact with users through graphical icons and visual indicators
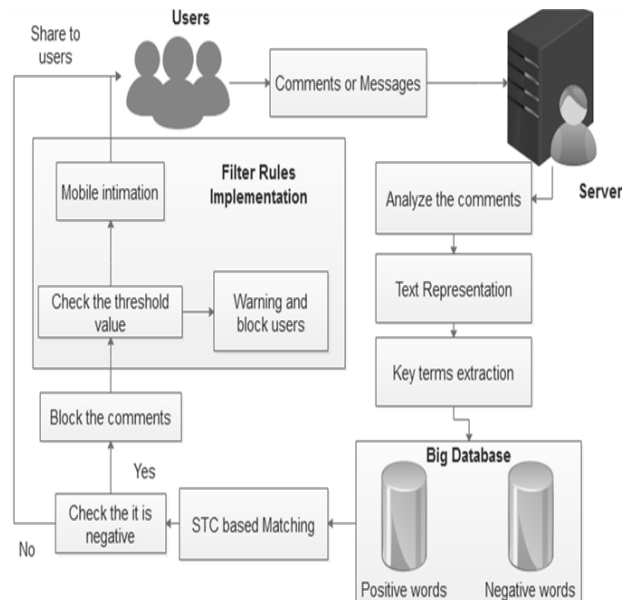
Fig 3: Representation of Proposed System

Post comments:

- This module is used to get the input from social users
- Comments may be various forms such as links or texts or short texts
- Comments are read and send to server page

STC implementation:

- In this module, comments are collected and stored in database
- Represent the text and splitting into phrases
- Pick the keywords and check whether it is positive or negative

Filtered GUI:

- Rules are generated by server for setting threshold values
- Based on threshold values, system can block friends who has posted negative comments
- A mobile intimation for blocking a friend has been issued

5.2 Blacklist

The major implantation is the execution of blacklist mechanism in order to keep away message from undesired user or creators. Blacklist ishandled undeviated by the system and this decides the user to be inserted into blacklist by using the predefined threshold value. It also decides the user prevention in the blacklist, it gets over. There are some list of rules to enhance the stiffness, such as owner can identify which user should be blocked with respect to the relationship in Online Social Network(OSN) and the user's id.

The user also have negative point of view about the user who is banned for a time period. We have a collection of data based on negative behaviour of the user. There are two assumptions;one is that the user will be inserted in blacklist for N number of times with the time period.

The person must be worthy for staging in blacklist for sometimes. The user who is inserted in blacklist at least once will apply this principle. Relative frequency is the one which is used to identify the system, whose information continued to fail the Frequency Relative (FR).Two constrains that can be used to measuredglobally and locally, which will consider the message by global; it will acknowledge all the user wall.

V. CONCLUSION

In this paper, the problem faced on posting negative comments on user wall has been controlled. The STC system classifies and labels the texts based on their polarity in the document. Threshold value for blacklisting a person is based on the number of negative comments they post. This work intent to capture negative comment and suggests blacklisting the holder of the comment. This paper can be further enhanced for identifying polarity ofmultimodal data, since opinions can be posted in any modes of data such as, text, image, video, etc.,

REFERENCES

1. B. Wang, G. Chen, L. Fu, L. Song, and X. Wang, "Drimux: Dynamic rumor influence minimization with user experience in social networks," in Proc. 30th AAAI Int. Conf. Artif. Intell. Feb. 2016.
2. D. N. Yang, H. J. Hung, W. C. Lee, and W. Chen, "Maximizing acceptance probability for active friending in online social networks," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2013, pp. 713–721.
3. S.Palanivel Rajan, C.Vivek, "Analysis and design of microstrip patch antenna for radar communication", Journal of Electrical Engineering & Technology, Online ISSN No.: 2093-7423, Print ISSN No.: 1975-0102, Volume 14, Issue 2, pp 923–929, DOI : https://doi.org/10.1007/s42835-018-00072-y, 2019.
4. L. Fu, W. Huang, X. Gan, F. Yang, and X. Wang, "Capacity of wireless networks with social characteristics," IEEE Trans. Wireless Communication., vol. 15, pp. 1505–1516, Feb. 2016.
5. S.Palanivel Rajan, K.Sheik Davood, "Performance Evaluation on Automatic Follicles Detection in the Ovary", International Journal of Applied Engineering Research, ISSN No.: 0973- 4562, Vol. 10, Issue 55, pp. 1-5, 2015
6. A. Montanari and A. Saberi, "The spread of innovations in social networks," in Proc. National Academy of Sciences of the United States of America PNAS, Aug. 2010, pp. 20 196–20 201
7. S.Palanivel Rajan, V.Kavitha, "Diagnosis of Cardiovascular Diseases using Retinal Images through Vessel Segmentation Graph", Current Medical Imaging Reviews (Bentham Science Publisher), Online ISSN No.: 1875-6603, Print ISSN No.: 1573-4056, Vol. No.: 13, Issue : 4, pp. 454-459, (Impact Factor–0.613), 2017
8. X. Rong and Q. Mei, "Diffusion of innovations revisited: From social network to innovation network," in Proc. 22Nd ACM Int. Conf. Inf. Knowledge Management., 2013, pp. 499–508.
9. C. Budak, D. Agrawal, and A. E. Abbadi, "Limiting the spread of misinformation in social networks," in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 665–674.
10. A. Bessi, F. Petroni, M. Del Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Viral misinformation: The role of homophily and polarization," in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 355–356.
11. E. Serrano, C. A. Iglesias, and M. Garijo, "A novel agent-based rumor spreading model in twitter," in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 811–814.
12. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2003, pp. 1175–1180.
13. C.Vivek, S.Palanivel Rajan, "Z-TCAM : An Efficient Memory Architecture Based TCAM", Asian Journal of Information Technology, ISSN No.: 1682-3915, Vol. No.: 15, Issue : 3, pp. 448-454, 2016
14. S.Vijayprasath, S.Palanivel Rajan, "Performance Investigation of an Implicit Instrumentation Tool for Deadened Patients Using Common Eye Developments as a Paradigm", International Journal of Applied Engineering Research, ISSN No.: 0973-4562, Vol. 10, Issue No.1, pp. 925-929, 2015.