Analytical and Technical Survey on Social Networking Sites

Karuna C. Gull
Dept. of Computer Science
KLE Institute of Technology,
Hubballi, Karnataka, India

Dr. Rajasekaran C.
Chaudhary Charan Singh University,
Meerut, Uttar Pradesh, India

Seema C. G
Dept. of Computer Science
Rani Channamma University
Belagavi, India

Suvarna G Kanakaraddi
School of Computers
KLE Technological University
Hubballi, Karnataka, India

***Abstract* — Online social media may be a paradigm that deals with production, transfer and consumption of data. User generated content within the kind of posts, comments, likes, and tweets establish a association between the producers and also the shoppers. Content selling and social media create a good team. Social media strategy is needed to finish the content selling method. Obtaining the contents from social media is extremely abundant vital in constructing a model that may do correct prediction. The prediction results change firms to realize feedback and insight a way to improve and market the product. They will also think about their promotion in higher sense. To hold out these varieties of analysis one has got to do deep survey on social media or social networking sites to gather massive scale of information. Then one has got to determine what quite tools is accustomed do the analysis of information collected. Thus to seek out the correct findings, the paper concentrates on a review of social media or social networking sites and varied tools that may be used for identical.**

Keywords — Social Networking Sites (SNS), Knowledge Discovery and Data mining (KDD), Support Vector for Regression (SVR), Support Vector Machine (SVM), K-Nearest Neighbour(KNN).

I. INTRODUCTION

Social structure forms a social network if people square measure connected directly or indirectly to every alternative supported common interests [3]. To know the structure and behaviour of social networks, one has to do analysis study of social networks.

Popularity of Facebook or Twitter as social networks square measure increasing day by day as they are providing sensible numbers of opportunities [5]. A decent survey is mentioned by authors [5] regarding social networking sites like quality of Facebook in Nov 2009, with over 316 million users, Turkey children square measure within the third place within the usage of Facebook, Turkey has fourteen million Facebook members. SNS given chance for creation of internet based web applications. At same time, opinions of the scholars and their associations, any existing members in SNS, square measure investigated by using the association rules technique.

According to Poll twenty two percent of teenagers go surfing to their favourite social media web site quite ten times on a daily basis. Quite half adolescents go surfing to a social media web site quite once a daily basis. Seventy five percent of teenagers currently own cell phones [6].

Impact of Social media is extremely high on our culture, in business, on the world-at-large. a number of them cause each positive and negative impact on the lifetime of people.

Kids and Adolescents: Enhance their individual and collective power through development and sharing. Their ideas square measure improved from creation of blogs, videos etc. They are open to many opportunities for community engagement. One's individual identity and unique social skills are fostered.

Enhanced Learning Opportunities: Students join themselves or attach with each other to social media to carry out the team assignment or projects to satisfy outcomes. Some colleges use blogs of those as teaching tools.

Power in Business: Communication between folks and businesses brings a lot of support to the business. It is a decent and effective approach which supports and suggests to boost income, image and recognition business organisers.

Power on Politics: Social websites have compete a vital role in several elections round the world, together with the countries United States, Iran, and India. Social media has conjointly served to rally folks for a cause and has galvanized mass movements in several countries.

Other Powers: To re-connect with their old friends and acquaintances or to have new friends, share content and footage. Keep updated of the most recent world and native developments and find out about completely different cultures and societies.

There square measure sure cons of social media too. Several introverts and socially reclusive users place an excessive amount of stress on virtual interaction, and ignore the real world outside. Cyber Bullying and online Harassment, influence of Advertisements on shopping for. Folks do not speak to their oldsters on the phone for a protracted time any more they like statuses, share articles with one another and chat regarding trifles of the day. Uploading videos on YouTube, passing desecrated pictures through Instagram and influencing teenagers through Twitter and Facebook, making WhatsApp cluster square measure the recent activities done by them [6] currently.

A. Why chosen Facebook as a case study?

Facebook is defined as "a social utility that helps people share information and communicate more efficiently with their friends, family and co-workers" (facebook.com)[15].The mission of Facebook is "Giving people the facility to share & make the globe open & connected". As of November 2009, with 316 million users, Facebook is that the most well liked and well-known social network throughout the globe. Facebook is currently the second most well liked site within the world just after search site Google per Alexa traffic rankings.

B. Why chosen Twitter as a case study?

Social media analysis may be a new epoch in marketing research exploiting the facility of real-time big data and state-of-the-art data analytics. Twitter popularity is growing, and also the most interesting aspect from the information analysis point of view, is that an oversized quantity of knowledge is publically available. As most of the people choose to publish their posts openly, in contrast to other social networks like Facebook or LinkedIn, where the data is simply accessible to those who are friends or connections. [8] Twitter has attracted many users by its exquisite features in an exceedingly short span. the advantages include Tweeting and Retweeting facility, Adding Favorites, Following people, Sharing, Can find the news, track the present Trends, Companies, Contacts, Celebrities and lots of more to pen. Thus an upscale source of knowledge is obtainable on twitter thanks to the increased usage. This rich data help in knowing users passion and excitement towards particular product. [9] Twitter users generate over 300M tweets every day, these users are overwhelmed by the large amount of knowledge available and also the huge number of individuals they'll interact with.

II. LITURATURE SURVEY

Measuring user influence of the more popular Social Networking site like Facebook or Twitter are challenging now. At the identical time activities done by users in SNS also matters for analysis. This survey paper [1] deals with the framework for predicting users influentiality in social networking site by taking Facebook as a case study. Authors depicts that the proposed work is completed by tracking the user information who use the allocation provided by them on SNS like Facebook

The paper [2] concentrates on the overall and security issues associated with the Social networking sites. Human interaction platforms like Facebook, Twitter, LinkedIn, Google, Yahoo etc show how they affect ones security and freedom. They discuss about the risks related to uploading sensitive information. during this paper, authors have talked about a number of the privacy and security concerns and prevention techniques that helps user to watch out while functioning on social media. Further the paper concentrates on the way to keep consumers safe while keeping their brand from being tarnished if an account is hacked or spoofed.

Authors [3] consider the study of the social networks, group formation within the social networks [4] and providing practitioners some useful concepts to web based applications' design for social network and lots of more. The literature focuses on the articles which emphasizes on centrality, linkage potency, identity, belief, activity and benefits. Thus the objective of paper is providing useful source of knowledge or knowledge to researchers who add and out of the social networking area.

This study[5] discovers the access frequency and usage time of Facebook by using various decision tree algorithms, ANN and SVM. Prediction functionality is formed more precise by using more training data. A priori algorithm will be seen useful to urge results, for the communication between classmates is over to communication between students and teachers just in case of Facebook. Students opinion about SNS is that it has been the most effective media for the production of resources.

III. ANAYTICAL PROCEDURES FOR DATA FROM SOCIAL SITES

A. What is not Data or knowledge Mining?

- Look up number in phone directory
- Query an Google search engine for data concerning "Amazon"

B. What is Data or knowledge Mining?

It has many Definitions:

Data Mining is "Uncover Hidden Information" / "knowledge discovery and data mining or processing (KDD)"

Data mining is "extraction of helpful patterns from knowledge sources, like databases, texts, web, image". Patterns should be: valid, novel, probably helpful, comprehensible.

Example of discovered patterns with Association rules is "80% of consumers who purchase cheese and milk also purchase bread, and 5% of customers buy all of them together" Cheese, Milk→ Bread [sup =5%, confid=80%].

C. Origins of Data Mining

Fig 1 draws ideas from machine learning / AI, pattern recognition, statistics, and database systems. Traditional Techniques may be unsuitable by reason of Enormity of knowledge. High spatiality of knowledge and Heterogeneous, distributed nature of knowledge

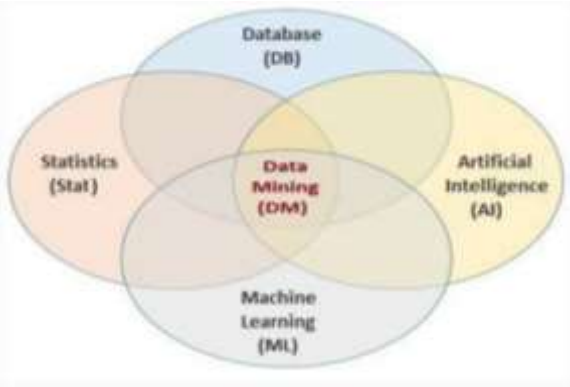

Fig. 1. Origin of Data Mining [Courtesy : https://pbs.twimg.com/media/DPsEcZaUEAA3VaS.jpg:large]

D. Machine Learning

Data Mining, [10] "Uses several machine learning techniques". Machine Learning: area of AI that examines the way to write programs that may learn. It is typically employed in classification and prediction tasks of data mining. Machine learning typically deals with static datasets. A variety of learning based on algorithms or tools are expressed downside:

Supervised Learning: Learns without knowledge of correct answers and learns by example. In supervised learning the classes are known in advance and also their types.

Here the output (dependent variable) depends on the input variable (independent variable). In some set of given supervisions the responder tries to calculate the required objective[11]. Two broad categories: Regression and Classification.

1) Regression Models

Regression models (both linear and non-linear) are used for predicting a true value, like remuneration for instance. If our independent variable is time, then we are forecasting future values, otherwise our model is predicting present but unknown values. Regression techniques vary from Linear Regression to SVR and Random Forests Regression.

Here we will perceive and suppose the sequence within which one will proceed with study of Machine Learning Regression models

1. Simple Linear Regression
2. Multiple Linear Regression
3. Polynomial Regression
4. Support Vector for Regression (SVR)
5. Decision Tree Classification
6. Random Forest Classification

After learning concerning these six regression models, one can have most likely the subsequent queries in their minds like however can we apprehend that model to settle on for our problem?, however can we able to improve every of those models? and lots of.

Let us try and answer these questions:

Before that we would like to seek out the pros and cons of every model. Then proceed to seek out answers.

How do we know that model to settle on for our problem?

- First, we need to figure out whether our problem is linear or non linear.
- Then, if our problem is linear, we should go for Simple Linear Regression if we only have one feature and Multiple Linear Regression if we have several features.
- If our problem is non linear, we should go for Polynomial Regression, SVR, Decision Tree or Random Forest.
- Then which one should we choose among these four? The method consists of using a very relevant technique that evaluates our models performance, called k-Fold Cross Validation, and then picking the model that shows the best results.

How can we improve each of these models?

- Parameter Tuning will allow us to improve the performance of our models.
- We need to study various parameters that support tuning in the models preferred to improve the performance.

2) Classification Models

Unlike regression where we predict a continuous number, we use classification to predict a category. There is a wide variety of classification applications from medicine to marketing. Classification models include linear models like Logistic Regression, SVM, and nonlinear ones like K-NN, Kernel SVM and Random Forests. In this part[12], we will understand and learn the way we are going to proceed with study of the subsequent Machine Learning classification models

- Logistic Regression
- K-Nearest Neighbours (K-NN)
- Support Vector Machine (SVM)
- Kernel SVM
- Naive Bayes
- Decision Tree Classification
- Random Forest Classification

After learning regarding these seven classification models, one can have most likely the subsequent queries in their minds like however will we grasp that model to decide on for our problem?, however can we able to improve each of those models? and plenty of additional.

Let us attempt to answer these questions:

Before that we'd like to search out the pros and cons of every model. Then proceed to search out answers.

How do we do know that model to decide on for our problem?

Same as for regression models, we tend to 1st got to comprehend whether or not our downside is linear or non linear. Then: If our downside is linear, we should always opt for Logistic Regression or SVM.

- If our downside is non linear, we should always opt for K-NN, Naive Bayes, Decision Tree or Random Forest.
- Then that one ought to we decide in every case? Model choice with k-Fold Cross Validation.
- Then from a business purpose, we might rather use:
  - Logistic Regression or Naive Bayes after we wish to rank our predictions by their likelihood.
  - For example if we would like to rank our customers from the very best likelihood that they purchase a precise product, to all-time low likelihood.
  - Eventually that permits we tend to focus on our promoting campaigns.
  - And in fact for this kind of business downside, we should always use Logistic Regression if our downside is linear, and Naive Bayes if our downside is non linear.
  - SVM after we wish to predict to that section our customers belong to. Segments are often any reasonably segments, as an example some market segments we tend to known earlier with clustering.
  - Decision Tree after we wish to possess clear interpretation of our model results.
  - Random Forest after we area unit simply trying to find high performance with less would like for interpretation.

How can we able to improve every of those models?

- Same answer as in Regression

Unsupervised Learning: learns without information of correct answers. In Unsupervised learning classes don't seem to be already celebrated, and also the learning method tries to search out acceptable classes. Here there's no superintendence thus system tries to adapt itself to things and learns manually supported some measure[13]. One broad category: clustering.

1) Clustering Models

Clustering is similar to classification, but the basis is different. In Clustering we don't know what we are looking for, and we are trying to identify some segments or clusters in our data. When we use clustering algorithms on our dataset, unexpected things can suddenly pop up like structures, clusters and groupings we would have never thought of otherwise. During this half, we are going to perceive and learn the way we are going to proceed with the study of the subsequent Machine Learning clustering models:

1. K-Means Clustering
2. Hierarchical Clustering

After learning regarding these two clustering models, one can have most likely got to grasp the execs and cons of those clustering techniques. The execs and cons of each are narrated as:

K-Means - execs as straightforward to grasp, simply labile, works well on little or giant datasets, fast, economical and performant and cons as got to opt for the amount of clusters *Hierarchical Clustering*- execs because the best range of clusters are often obtained by the model itself, sensible image with the dendrogram and cons as Not acceptable for big datasets

Natural language process (NLP) may be a field of applied science, computer science, and linguistics involved with the interactions between computers and human (natural) languages. As such, NLP[7] is expounded to the realm of human–computer interaction. several challenges in NLP involve: tongue understanding, facultative computers to derive that means from human or tongue input; et al involve tongue generation. fashionable NLP algorithms area unit supported machine learning, particularly applied mathematics machine learning. The paradigm of machine learning is totally different from that of most previous tries at language process. previous implementations of language-processing tasks generally concerned the direct hand committal to writing of huge sets of rules. The machine-learning paradigm calls

instead for victimization general learning algorithms — typically, though not continually, grounded in applied mathematics logical thinking — to mechanically learn such rules through the analysis of huge corpora of typical real-world examples. A corpus (plural, "corpora") may be a set of documents (or generally, individual sentences) that are hand-annotated with the right values to be learned. many various categories of machine learning algorithms are applied to NLP tasks[14]. These algorithms take as input an outsized set of "features" that area unit generated from the input file. a number of the earliest-used algorithms, like call trees, created systems of arduous if-then rules like the systems of hand-written rules that were then common. more and more, however, analysis has centered on applied mathematics models, that build soft, probabilistic selections supported attaching real-valued weights to every input feature. Such models have the advantage that they will categorical the relative certainty of the many totally different doable answers instead of just one, manufacturing additional reliable results once such a model is enclosed as a part of a bigger system

Weka Tool: The goal of Weka is to build a state-of-the art facility for developing machine learning (ML) techniques and allow people to apply them to real-world data mining problems[16].

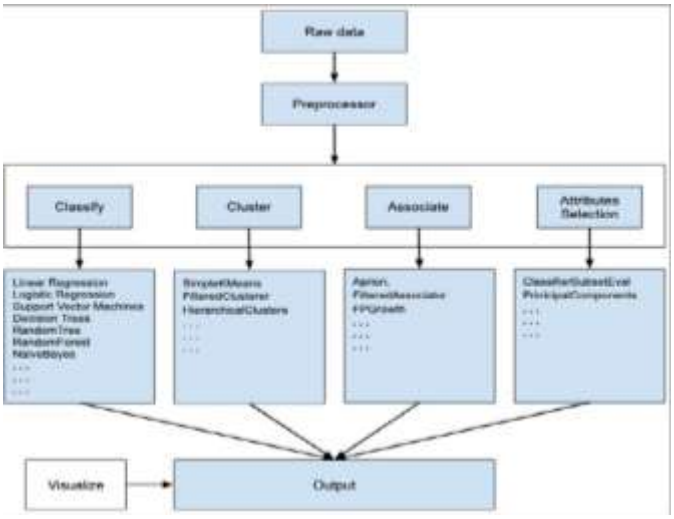

Fig. 2. Features of Wek Tool

Weka helps in building the state-of-the-art (Brand new) facility for developing techniques for Machine Learning and investing their applications in the key areas of ML as shown fig.2. Weka provides the workbench for ML and Data Mining. It determines the factor that contributes towards its successful application in agriculture, industries, scientific research and developing new methods for ML and ways of accessing their effectiveness [17].

The sample working with Weka tool is expressed using Explorer, Experimenter, Knowledge Flow and Simple CLI as depicted in Fig.3 to Fig.7.

Fig. 3. Main Screen of Weka Tool

Fig. 4. Preprocessing of Data

Fig. 5. Setting up Experiment Environment

Fig. 6. Flow of Dataset within Experiment Environment for visualising results

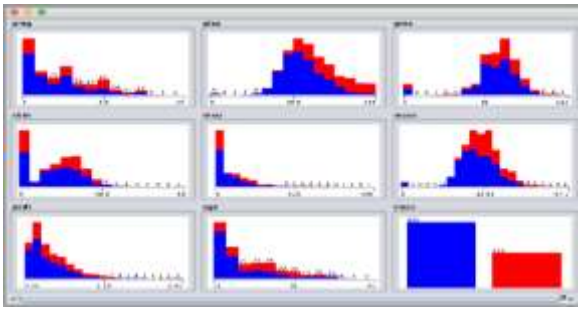

Fig. 7. Visualization of Results of ML techniques

Feature of Weka Tool are User Friendly, Poratble, Easy To Use Gui And Command Line and Provides Access To Sql Database.

IV. CONCLUSION

It is up to every user to use social sites sagely to boost their skilled and social life and exercise caution to make sure that they are doing not fall victim to on-line dangers. At the end, if we will keep our own life focused in point of fact and use social networking as a tiny low a part of it, we should always be simply fine. For those that can't, it'd be time to show off the PC for a small amount and opt for a walk. Since it's been a survey relating to the activities of social sites, with specification why Facebook and twitter as case studies, it deals with execs and cons of these sites. Moving further literature review part of the paper talks regarding SNS, numerous techniques to gather, preprocess the information. Finally as a concluding part of the survey paper gives glimpse on various machine learning

techniques that can be used to carry out the analysis of the preprocessed data

References

[1] Shwetha Bhat, Karuna Gull, Akshata Angadi, (2013) "Framework for Identifying Influential User in Online Social Networking Site" International Journal of Latest Trends in Engineering and Technology (IJLTET). Special Issue – IDEAS-2013. Pp.97-106 ISSN: 2278-621X.

[2] Kiran B. Malagi, Akshata Angadi, Karuna Gull (2013) "A Survey on Security Issues and Concerns to Social Networks", International Journal of Science and Research (IJSR), Volume 2, Issue 5, May, 2013. pp.256-265. ISSN: 2319-7064.

[3] Karuna C. Gull, Padmashri Desai, Akshata B. Angadi, (2012) "Theoretical Concepts of Social Networks and Group Formation: A Survey". International Journal of Engineering Innovation & Research (IJEIR), Volume 1, Issue 6, Nov-Dec, 2012. pp.577-583. ISSN: 2277-5668.

[4] Granovetter, M. S. (1973) The strength of weak ties. American *Journal of Psychology*, 78 (6), pp.1360-1380.

[5] Karuna C. Gull, Dr. Santosh kumar Gandhi, Santoshkumar B. Shali, (2013) "Application of Data mining Technique in Social Networking Sites: Facebook as a Case", International Journal of Advanced Research in Computer Science (IJARCS), Volume 4, No. 2, Jan-Feb, 2013. pp.320-326. ISSN: 0976-5697.

[6] https://www.slideshare.net/KunalGawade2/social-media-power-positive-or-negative

[7] Michael Speriosu, Nikita Sudan , Sid Upadhyay and Jason Baldridge ,"Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph", Proceedings of EMNLP,Conference on Empirical Methods in Natural Language Processing,2011, pp. 53-63.

[8] Albert Bifet , Geoff Holmes and Bernhard Pfahringer "Detecting Sentiment Change in Twitter Streaming Data", 2nd Workshop on Applications of Pattern Analysis, JMLR : Workshop and Conference Proceedings 17, 5-11, 2011

[9] Tim Trampedach, "Introduction to Social Networking" in rockyou

[10] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2000. Available: www.mkp.com/datamining.

[11] https://www.udemy.com/course/machinelearning/learn/lecture/9236710#overview

[12] https://www.superdatascience.com/sds-041-inspiring-journey-totally-different-background-data-science/

[13] http://www.superdatascience.com/2

[14] Bo Pang, and Lillian Lee, "Thumbs up?: sentiment classification using machine learning techniques, EMNLP '02", the ACL-02 conference on Empirical methods in natural language processing - Volume 10, Pages 79-86, Association for Computational Linguistics Stroudsburg, PA, USA ©2002

[15] Karuna C. Gull, Dr. Santosh kumar Gandhi, Santoshkumar B. Shali, (2013) "Application of Datamining Technique in Social Networking Sites: Facebook as a Case", International Journal of Advanced Research in Computer Science (IJARCS), Volume 4, No. 2, Jan-Feb, 2013. pp.320-326. ISSN: 0976-5697.

[16] Witten E. Frank, Datya Mining, Practical Machine Learning Tools and Tecniques with JavaImplementation, Morgan Kaufmann Publishers, 2000

[17] R.Kirkby, WEKA Explorer User Guide for version3.3-4, University of Weikato,2002