Mangalayatan Journal of
Scientific and Industrial Research

Man. J. Sci. Ind. Res.
Volume 1 (1) (2024), 24106

Article

On Problems in social media and their Solution by Using Techniques of Machine Learning, Artificial Intelligent and Data Mining

Anjali Bhardwaj

Dept. of statistics, Institute of social sciences, Dr. Bhimrao Ambedkar university, Agra
Corresponding author: dranjalibharwaj0506@gmail.com

Abstract

This paper provides idea of data mining including evolution of data mining, data mining parameters, data mining Process, Architecture of data mining system, types of data mining system, data mining algorithms and techniques. All these techniques are having their own merits and demerits. It focuses on social media mining which is the core of this paper. This paper discusses the most frequently used social media mining techniques such as SVM, BN and DT. Due to uniqueness of social media data – velocity, size, dynamism, noisy, unstructured, heterogeneous behavior.etc, researchers are invited to do more research on existing and upcoming technologies. Hopefully in future work there will be further explored in data mining algorithms, including their impact and new research issues.

Keywords: social media, Machine learning, social network analysis, Applications of social media analysis

1. Introduction

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data preprocessing, model and inference considerations, interestingness, metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets. Data Mining is all about discovering unsuspected/ previously unknown relationships amongst the data.

Originated from knowledge discovery from databases (KDD), also known as data Mining (DM), Data Mining (DM) is the extraction of new knowledge from large databases. Many techniques are currently used in this fast-emerging field, including statistical analysis and machine learning based approaches. With the rapid development of the World Wide Web and the fast increase of unstructured databases, new technologies and applications are continuously coming forth it his field.

Data Mining is referred so as Information Harvesting / Knowledge Mining / Knowledge Discovery in Databases / Data Dredging / Data Pattern Processing / Data Archaeology / Database Mining, Knowledge Extraction and Software. Data Mining is a process of analyzing data from many different dimensions or angles and summarizing it into useful information that can be applied in different fields to take proper decision. It increases profits and cuts costs, or both. Technically, data mining is the computing process of discovering patterns or correlations in large relational databases involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

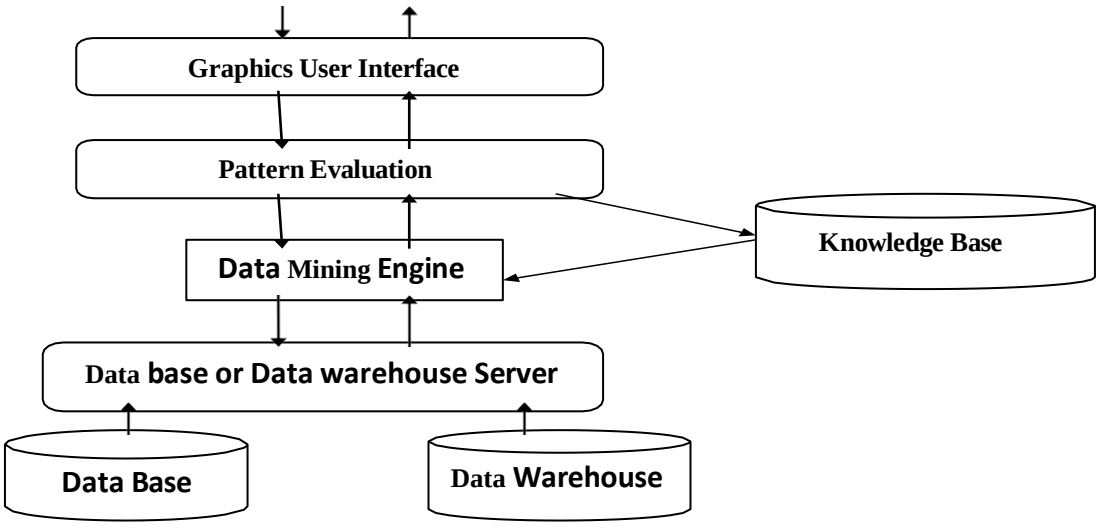


Figure 1: KDD Process

Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

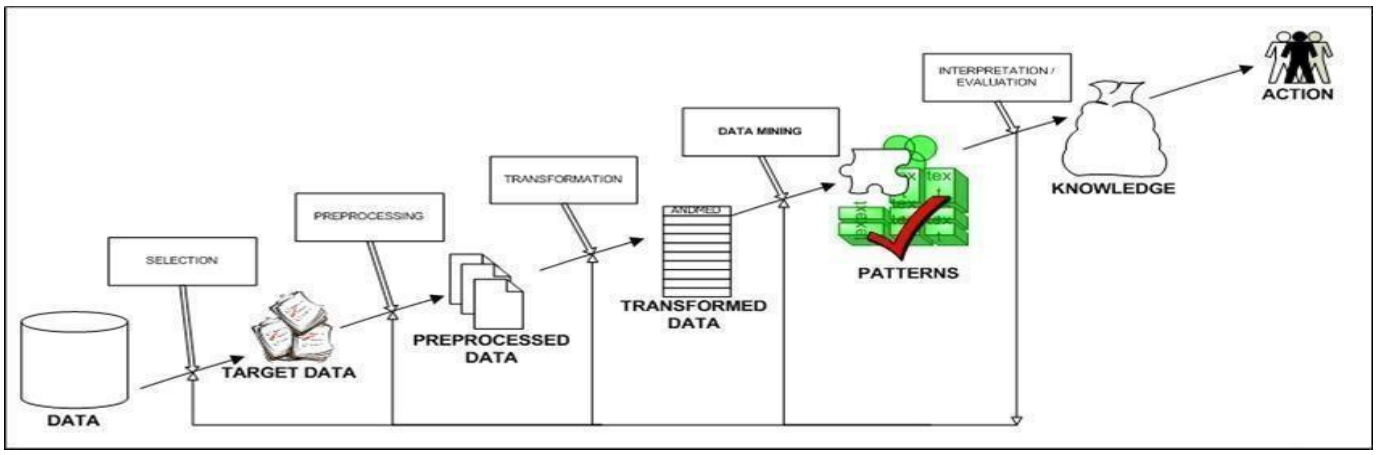


Figure 2: The most commonly used techniques in data mining are:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.

Rule induction: The extraction of useful if-then rules from data based on statistical significance.

Social media are interactive computer-mediated technologies that facilitate the creation or sharing of information, ideas, career interests and other forms of expression via virtual communities and networks. The variety of stand-alone and built-in social media services currently available introduces challenges of definition; however, there are some common features: Social media are interactive Web 2.0 Internet-based applications.

1. User-generated content, such as text posts or comments, digital photos or videos, and data generated through all online interactions, is the lifeblood of social media.
2. Users create service-specific profiles and identities for the website or app

that are designed and maintained by the social media organization.

3. Social media facilitate the development of online social networks by connecting a user's profile with those of other individuals or groups.

Users usually access social media services via web-based technologies on desktops and laptops, or download services that offer social media functionality to their mobile devices (e.g., smartphones and tablets). As users engage with these electronic services, they create highly interactive platforms through which individuals, communities, and organizations can share, co- create, discuss, participate and modify user-generated content or self-curated content posted online.

Networks formed through social media change the way groups of people interact and communicate or stand with the votes. They "introduce substantial and pervasive changes to communication between organizations, communities, and individuals. These changes are the focus of the emerging fields of techno self studies. Social media differ from paper-based media (e.g., magazines and newspapers) and traditional electronic media such as TV broadcasting, Radio broadcasting in many ways, including quality, reach, frequency, interactivity, usability, immediacy, and performance. Social media outlets operate in a dialogic transmission system (many sources to many receivers). This is in contrast to traditional media which operates under a mono-logic transmission model (one source to many receivers), such as a newspaper which is delivered to many subscribers, or a radio station which broadcasts the same programs to an entire city. Some of the most popular social media websites, with over 100 million registered users,include Facebook, YouTube, WeChat, Instagram, QQ, QZone, Weibo, Twitter, Tumblr, Telegram, Baidu Tieba, LinkedIn, WhatsApp, LINE, Snapchat, Pinterest, Viber, VK, Reddit, and more.

Observers have noted a wide range of positive and negative impacts of social media use. Social media can help to improve an individual's sense of connectedness with real or online communities and can be an effective communication (or marketing) tool for corporations, entrepreneurs, non-profit organizations, advocacy groups, political parties, and governments.

2. Data mining in social media

Social media data mining is on the rise. The increasing availability of data on users and their online behaviour, the decreasing cost of collecting, storing and processing data, and the exponential expansion of social media platforms from which much of this data is taken mean that – at least in theory – an increasingly diverse range of actors can mine social data. This process can involve simply counting the likes and shares of social media content, or more advanced analysis of its strength, sentiment, passion, reach and other quantifiable characteristics (mentions, users, sources, hashtags).

The metadata that sits behind social media content is also widely mined, and considered by some to be more valuable than the content itself.

Social media data mining is under taken by the major platforms themselves (like Facebook and Twitter), by intermediary commercial companies, or with tools which are free to all comers; some easy-to-use (for example Social Mention), and others more complex. Methods for analysing social media data promise powerful new ways of knowing publics and capturing what they say and do. And yet access to these methods is uneven, with large corporations and governments tending to have the best access to data and analytics tools. Critics warn of a number of troubling consequences for publics that result from the rise and spread of data mining: less privacy, more surveillance and social discrimination, and a new means of controlling how publics come to be represented and so understood. Meanwhile, the tools and systems that generate knowledge from social media data are typically opaque and are rarely open to public scrutiny and supervision. We argue that these various enactments and characteristics of data mining are constitutive of a new form of data power.

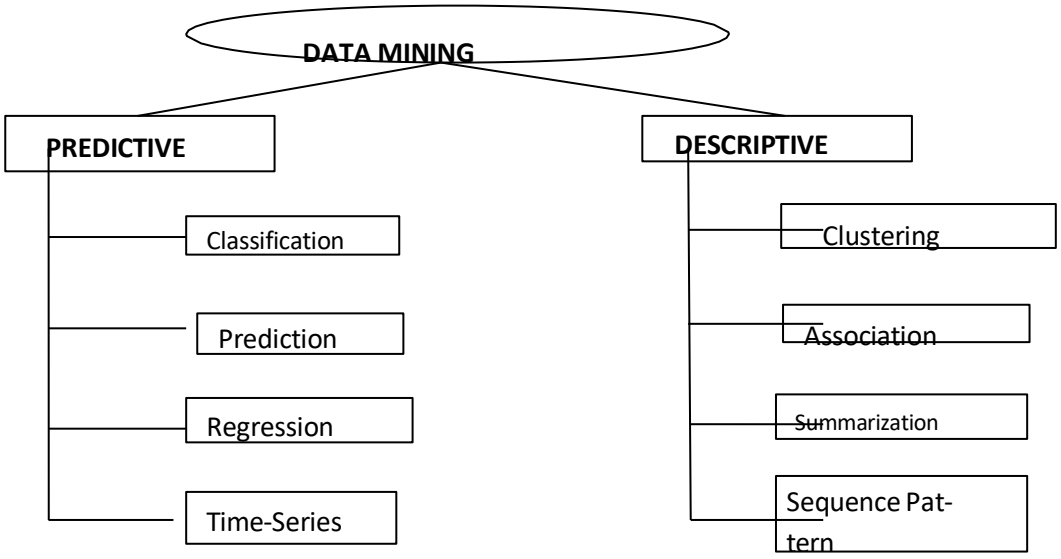
Roots of data mining

1. **Statistic-** The most important lines is statistics. Without statistics, there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analyses are underpinned. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role.

2. **Artificial intelligence and machine learning-** Data mining's second longest family line is artificial intelligence and machine learning. AI is built upon heuristics as opposed to statistics, and attempts to apply human-thought like processing to statistical problems. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at thievery high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. Machine Learning could be considered as an evolution of AI, because it blends AI heuristics with advanced statistical methods. It let computer programs learn about the data they study and then apply

learned knowledge to data.

Data mining technique



3. Data mining technique used in social media

An algorithm in data mining or machine learning creates a model by analyzing input data and extracting specific types of patterns or correlations. This output is analyzed over many iterations to find the optimal parameters for creating model. These parameters are then applied across the entire data sets to extract actionable patterns. Choosing the best algorithm for a specific analytical task can be a challenge. Different algorithms can be used to perform a same task, each algorithm produces different results, and some algorithms can produce more than one type of result. More number of algorithms are available in the field to fulfill the needs of users. Some of the top most algorithms are: 1) Support vector machines (SVM), 2) Bayesian Networks (BN), 3) Decision Tree (DT), 4) C4.5, 5) K-Nearest Neighbors (KNN) 6) k-means, 7) Apriori, 8) Expectation-Maximization (EM), 9) Page Rank, 10) AdaBoost, 11) Naive Bayes, 12) CART

Below mentioned issues are frequently observed in social media data. Data velocity or incoming of data is enormous and dynamic nature of data is also unpredictable. The massive data available in social media is unstructured and heavily used by youngsters. So, it is a heavy duty imposed on researchers to find out suitable algorithms to fetch correct patterns / trends / correlations existing.

among data which are helpful to youngsters for their bright future as well as the nation. 19 data

mining techniques had been applied by researchers in the area of social media. Among them SVM, BN, and DT are the **most applied techniques in the area of social media.**

1. **Support Vector Machines (SVM)**

   Support vector machine (SVM) classify data into 2 classes using the concept of a hyper plane. SVM is similar to C4.5 except SVM doesn't use decision trees at all. For a simple classification task with just 2 features, the hyper plane can be a line,

   $$y = m x + b.$$

   For example: Take a bunch of balls of red and blue colors on a table which aren't too mixed together. In this case they can be separated with the stick. When a new ball is added on the table, by knowing which side of the stick the ball is on, its color can be predicted. Here the balls represent data points, and the red and blue colors represent 2 classes. The stick represents the simplest hyper plane which is a line. This is a supervised learning, since a dataset is used to first teach the SVM about the classes. Only then is the SVM capable of classifying new data. SVM along with C4.5 are generally the classifier. No classifier will be the best in all cases. Kernel selection and interpretability are some weaknesses of SVM.

   Interpretability means representation of acquired knowledge in a readable form. In Decision trees, it is good. But not in KNN, SVM.

2. **Bayesian Networks (BN)**

   Bayesian (also called Belief) Networks (BN) is a fundamentally important DM technique consisting of two parts, the directed acyclic graph G, with nodes (attributes) and arcs (direct dependencies) and the conditional probability tables for each node.

   Bayes classifier achieves the optimal result by applying the probability theory. BN represent events and causal relationships between them as conditional probabilities involving random variables. Given the values of a subset of these variables (evidence variables), BN computes the probabilities of another subset of variables (query variables) However, the Bayesian approaches can not cross the need of the probability estimation from the training dataset. It is noticeable that in some situations, where the decision is clearly high degree of randomality, the Bayesian approaches will not be a good choice.

3. Decision Tree(DT)

Decision tree is a classifier that constructs tree structure with nodes and arcs. Root and internal nodes are labeled with question. Arc represents the answer to the associated question. Each leaf node indicates a prediction of a solution to the problem / value of target variable. Decision tree predicts information in the form of rules which are if-then- else expressions. This outcome explains the decisions that lead to the prediction.

4. C4.5

C4.5 is a classifier in the form of decision tree that takes set of data and predicts which class the new data belongs to, based on labeled or training data that are classified already. For a given dataset of patients whose attributes are age, pulse, blood pressure, VO2max, family history, etc. C4.5 is loaded with these labeled / training data. When a new patient' record with attributes is given as input, C4.5 constructs a decision tree and predicts the class for new patient whether he will get cancer or not, based on their attributes.

This kind of learning is supervised learning, since the training dataset is labeled with classes. Based on this only, C4.5 forms a decision tree and predicts the class of new data. Difference between C4.5 and decision tree systems

- First, when C4.5 constructing decision tree it uses information gain.
- Second, C4.5 uses a single-pass pruning process to lessen over-fitting. As a result it produces many improvements
- Both continuous and discrete data are used in C4.5. It specifies ranges or thresholds for continuous data thus turning continuous data into discrete data.
- Finally, C4.5 handles incomplete data in its own ways.

4. K-Nearest Neighbors (KNN)

KNN is a classifier. It is a lazy learner. During training process it just stores the training data rather than doing something. When new unlabelled data is given as input, then only it does classification. C4.5, SVM are eager leaner who builds classification, clustering, segmentation, association rules, sequencing, decision tree various ALM reports like Statement of Structural Liquidity, Statement of Rate of Interest Sensitivity etc. or accounting reports like record and Profit & Loss Account are often generated instantaneously for any desired period/ date . Trends are often analyzed and predicted with the supply

of historical data and therefore the data warehouse assures that everybody is using an equivalent data at an equivalent level of extraction, which eliminates conflicting analytical results and arguments over the source and quality of data used for analysis. In short, data warehouse enables information science to be wiped out a reputable, efficient manner. The Committee recognizes the necessity for data warehouses and data processing both at the individual bank level and at industry level.

5. Literature review

Rahman, M. M, (2012) the authors have presented a systematical data mining architecture to mine intellectual knowledge from social data. In this research, they have used social networking site facebook as primary data source and collected different attributes like comments, me, wall post and age from facebook as raw data and used advanced data mining approaches to excavate intellectual knowledge and also analyzed their mined knowledge and suggested that Social data mining is an interesting and challenging research to mine intellectual knowledge which can be used in human behavior prediction, decision making, pattern recognition, social mapping, job responsibility distribution and product promoting.

Rahman, M. M, (2012), the author showcases a systematical data mining approach to mine intellectual knowledge from social data. The author took Facebook as a primary data source and proposes to use different data mining techniques to analyze this social networking site and other sites too. One algorithm that the author had discussed was K-nearest neighbor (K-NN). This algorithm classifies objects based on samples.

Mosley Jr, R. C. (2012), the author discusses the application of correlation, clustering, and association analyses to social media. The main purpose of this paper was to describe how data mining and text analytics can be applied to social media in order to identify key themes in the data. To be more specific the author described the analysis of Twitter posts. Certain issues in terms of accuracy while collecting the data from social media were also highlighted.

Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F (2014), the authors have studied the techniques that are currently used to analyze Social Media(SM). In this paper the analysis of SM data has been proved to be effective, this is so because of the capacity possessed by data mining in handling unstructured and dynamic data. According to the authors, in future to

mine the data generated on SM, research will be carried out on currently used and yet-to-be-explored data mining techniques.

XimingWang · Panos M. Pardalos, (2014), the authors have presented a survey on uncertainties of support vectors in SVM. In earlier methods, the values of data point / support vectors are known. Suppose if it is uncertain, SVM becomes more complex in classifying objects as well as in non linear kernel selection. Hence they have suggested that more research could be conducted to deal out this uncertainty of data points and selection of non linear kernel.

ThabitZatari. (2015) , the author has studied on the data mining techniques that are currently used to analyze Social Media data. In this paper, analysis has proved that it is unrealistic to expect one system to mine all kinds of data. Hence different kinds of data mining techniques are available in field for different applications. Every data mining algorithm / technique has its own strengths and limitations.

MohammadNoorInjadat, FadiSalo, Ali BouNassif (2016), the authors have summarized that social network data analysis, business and management were the most active domains that requiring mining of social media data and the most frequent social media mining techniques are SVM, BN, and DT. Also, the authors suggest that the area of social media still calls for more profound research to house a twin-focus method which incorporates accurate.

Pushpam, C. A., Jayanthi, J. G. (2017), the authors have studied that Knowledge plays a vital role in every sphere of human life. Data Mining supports to acquire knowledge by discovering pattern / correlations among data. This information is applied in various applications like business, education, social media, medical, Agriculture etc. Data mining field has attained enormous success from its inception to the present level. Also it faces many issues especially while handling social media data. Social media is one of the important sources that provide huge volume of data that are unstructured and heterogeneous. Handling this data is really a very big challenge to the researchers. At present, a number of data mining algorithms and techniques are available with their own merits and demerits. Finding a suitable algorithm for a particular application is a very big challenge. This paper imparts many issues in data mining and also focuses scope of the data mining in social media which will be helpful in the further research.

Balaji T.K., et al., (2021), the authors have summarized that social media (SM) are the most widespread and rapid data generation applications on the Internet increase the study of these

data. However, the efficient processing of such massive data is challenging, so we require a system that learns from these data, like machine learning. Machine learning methods make the systems to learn itself. Many papers are published on SM using machine learning approaches over the past few decades. In this paper, we provide a comprehensive survey of multiple applications of SM analysis using robust machine learning algorithms. Initially, we discuss a summary of machine learning algorithms, which are used in SM analysis. After that, we provide a detailed survey of machine learning approaches to SM analysis. Furthermore, we summarize the challenges and benefits of Machine Learning usages in SM analysis. Finally, we presented open issues and consequences in SM analysis for further research.

Observation and discussion

Support Vector Machine is very accurate, less over fitting and robust to noise. But interpretability is good in Decision trees, not in KNN and SVM. In Bayesian Networks, it cannot cross the need of the probability estimation, from the training dataset. It is noticeable that in some situations, where the decision is clearly based on certain criteria, or the dataset has high degree.

Conclusion

It can be concluded that social media mining is a new initiative to build new business strategies. The Social media houses vast amount of user-generated data which can be used for data mining, therefore guarantee a huge potential in terms of knowledge.

This review paper provides idea of data mining including evolution of data mining, data mining Process, Architecture of data mining system, types of data mining system, data mining algorithms and techniques. All these techniques are having their own merits and demerits.

Reference

1. AakankshaBhatnagar, Shweta P. Jadye, Madan Mohan Nagar (2012)"Data Mining Techniques & Distinct Applications: A Literature Review" in *International Journal of Engineering Research & Technology (IJERT)*, Vol. 1, Issue 9, PP:1-3.
2. Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F, (2014) "A Survey of Data Mining Techniques for Social Media Analysis" in *Journal of Data Mining & Digital Humanities*, PP:1-27.
3. Anmol Kumar1, Amit Kumar Tyagi2, Surendra Kumar Tyagi3 (2014), "Data Min-

ing: Various Issues and Challenges for Future :A Shortdiscussion on Data Mining issues for future work", in *International Journal of Emerging Technology and Advanced Engineering*, Volume 4, Special Issue 1, PP:1-8.

4. Annan Naidu Paidi (2012) "Data Mining: Future Trends and Applications" in *International Journal of Modern Engineering Research (IJMER)*, Vol.2, Issue.6, PP:4657-4663.
5. Babu N. V., Kanaga E. G. M., (2021). Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: *A Review. SN Computer Science.* (2022) 3:74
6. Balaji T.K., et al., (2021). Machine learning algorithms for social media analysis: A survey. *Computer science review.* 40.(2021)
7. B.I. Analytics(2014), Business Intelligence from Social Media A Study from the VAST Box Office Challenge, *IEEE*Comput. Graph. Appl. 34 58–69. doi:10.1109/MCG.2014.61.
8. B.J. Jansen, K. Sobel, G. Cook (2011), "Classifying ecommerce information sharing behaviour by youths on social networking sites", *J. Inf. Sci.* 37 (2011) 120–136. doi:10.1177/0165551510396975.
9. C.C. Yang, T.D. Ng(2011), "Analyzing and visualizing web opinion development and social interactions with density-based clustering", *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans*. 41 1144–1155. doi:10.1109/TSMCA.2011.2113334.
10. David Heckerman (2016), "Bayesian Networks for Data Mining and Knowledge Discovery", Volume 1, Issue 1.
11. Diaz-Garcia, J. A., Dolores Ruiz, M., Martin-Bautista, M. J. (2022). A survey on the use of association rules mining techniques in textual social media. *Artificial Intelligence Review* (2023) 56:1175–1200
12. Dinesh Bhardwaj1, Sunil Mahajan2 (2016), "Analysis Of Data Mining Trends, Applications, Benefits And Issues", in *International Journal of Computer Science and Communication Engineering*, Volume 5 issue 1, PP:53-57.
13. Dr.B.Umadevi1, P.Surya2 (2017), "A Review on Various Data Mining Techniques in Social Media", in *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 5, Issue 4, PP: 8082-8086.
14. Dr.M.Chidambaram, R.Umasundari (2016), "A Survey on Feature Selection in Data

Mining", in *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)* Volume4, Issue-1, PP: 13 -14.

15. Dr. PoonamChaudhary (2015) , "Data Mining System, Functionalities and Applications: A Radical Review" in *International Journal of Innovations in Engineering and Technology (IJIET),* Volume 5, Issue 2 ,PP:449-455.
16. Han, J. and M. Kamber (2001), "Data Mining: Concepts and Techniques", *Morgan Kaufmann.*
17. Han, J., M. Kamber, and A. K. H. Tung (2001), "Spatial Clustering Methods in Data Mining: A Survey", *H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery,* Taylor and Francis, 2001.
18. Heikki, Mannila (1996), "Data mining: machine learning, statistics, and databases", *Statistics and Scientific Data Management,* pp. 2-9.
19. HemlataSahu, " A Brief Overview on Data Mining Survey" in *International Journal of Computer Technology and Electronics Engineering (IJCTEE)* Volume 1, Issue 3, PP: 114-121
20. H. K. Chan1, E. Lacka2, R. W. Y. Yee3, M. K. Lim4 (2014), "A Case Study on Mining Social Media Data", in *the Proceedings of the 2014 IEEE IEEM, 978-1-4799-6410- 9/14/$31.00 © IEEE* , PP: 593- 596
21. Miller and J. Han (eds.) (2001),"Geographic Data Mining and Knowledge Discovery"*Taylor and Francis,*
22. Miss. Nazneentarannum S. H. Rizvi (2016), "A Systematic Overview On Data Mining: Concepts And Techniques" in *International Journal of Research in Computer & Information Technology (IJRCIT),* Vol. 1, Special Issue 1, PP:136-139.
23. MohammadNoorInjadat, FadiSalo, Ali BouNassif (2016) "Data Mining Techniques in Social Media: A Survey", NEUCOM17295, Volume 214, PP:654-670.
24. Mosley Jr, R. C. (2012), "Social media analytics: Data mining applied to insurance Twitter posts"*In Casualty Actuarial Society E-Forum,* vol 2 (p. 1)
25. Mrs. Bharati M. Ramageri, "Data Mining Techniques And Applications", in *Indian Journal of Computer Science and Engineering,* Vol. 1 Issue. 4, PP: 301-305.
26. M.S. Chen, J. Han, and P.S. Yu. (1999). "Data mining: An overview from database perspective", *IEEE Transactions on Knowledge and Data Eng.,* 8(6):866-883.

27. M. Zuber (2014). "A Survey of Data Mining Techniques for Social Network Analysis", *Int. J. Res. Comput. Eng. Electron*. 1–8.

28. Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2, and Rasmita Panigrahi3 (2012), "The Survey of Data Mining Applications And feature scope", in *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, Vol.2, Issue.3.

29. NesmaSettouti, Mohammed El Amine Bechar and Mohammed Amine Chikh (2016), "Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classification Task",in*International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 4, No.1, PP:46-51.

30. Nikita Jain, Vishal Srivastava (2013) "Data Mining Techniques: A Survey Paper" in *IJRET: International Journal of Research in Engineering and Technology*, Volume: 02, Issue: 11, PP:116-119.